Citation Chain Aggregation: An interaction model to support citation cycling

Timothy Cribbin
People and Interactivity Research Centre
Brunel University
Uxbridge, UK
timothy.cribbin@brunel.ac.uk

ABSTRACT
Citation chaining is a powerful means of exploring the academic literature. Starting from just one or two known relevant items, a naïve researcher can cycle backwards and forwards through the citation graph to generate a rich overview of key works, authors and journals relating to their topic. Whilst online citation indexes greatly facilitate this process, the size and complexity of the search space can rapidly escalate. In this paper, we propose a novel interaction model called citation chain aggregation (CCA). CCA employs a simple three-list view which highlights the overlaps that occur between the first-generation relations of known relevant items. As more relevant articles are identified, differences in the frequencies of citations made by or to unseen articles provide strong relevance feedback cues. The benefits of this technique are illustrated using a simple case study.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*Retrieval models, Relevance feedback, Search process*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries–*User issues.*

General Terms
Design, Human Factors

Keywords
Citation index, chaining, polyrepresentation, digital library, search user interface

1. INTRODUCTION
Chaining is a search mode that exploits the networks of relationships that emerge as a result of citation behavior. Backward chaining [4] or footnote chasing [1] involves looking up articles that are cited by a known seed article. In contrast, forward chaining [4] or citation searching [1] involves going forwards in time to find relevant articles that cite a known article. Cawkell [3] describes the method of *citation cycling*, in which backward and forward chaining are used in combination to build an overview of an unfamiliar field. The researcher begins by noting the most useful citations in some initial seed article. They then forward chain to the articles that cite these articles, note their citations and forward chain from these. As the cycle continues, key articles, authors and journals gradually become apparent.

Backward chaining is commonly applied by most academic disciplines, given the established practices of citation and referencing. In contrast, forward chaining has only recently been made feasible through the introduction of citation indexes. Since the 1990s these indexes have been accessible online through hypertext interfaces [3]. Services like Thomson-Reuter's Web of Science (WOS), ACM's Digital Library and the open-source CiteSeerX greatly facilitate locomotion through citation graphs by providing hyperlinks alongside each article record to cited and citing article records. Online citation indexes also enable other useful functionality such as the *related records* search, which ranks records according to the degree of bibliographic coupling with a seed article. Whilst there is little doubt that these visual interfaces have greatly facilitated the practice of citation cycling, two key problems still face the user engaged in citation cycling.

The first problem is one of managing the *size of the search space*. Even a single iteration of citation cycling can result in a large, heterogeneous search space, particularly on the forward chain side. A single influential article might cite perhaps a dozen or so well chosen works, but in turn be cited hundreds or even thousands of times, for a diversity of reasons. Without judicious filtering, the number of articles to follow-up can become unwieldy after just one or two iterations [3].

Deciding which articles to follow-up is a difficult task because authors' reasons for citation can vary widely in both purpose and salience [4]. Judging citations made by a known article is relatively simple because they reflect the intentions of a single project and are contextualized by their location within the text. However, the motivations for citations made to an article, reflecting the work of various authors over many years, may be quite diverse and not always immediately apparent from reading article surrogates [1]. Ranking search results by citation count alone is not always helpful if, for example, the researcher's interest does not relate to the primary impact of the seed. Citation counts also favor older articles over more recent publications. Related record searching provides a convenient proxy for citation cycling i.e. finding articles that cite some or all of the same articles as the seed. Whilst this can alleviate the latter issue, effective cycling still depends on the judicious selection of citations in order to control both the quality and quantity of accumulated articles [3, 6]. At the current time, none of the

citation indexes provides an option to refine a related search to only take selected citations into account.

The second problem can be best described as the *focus and context problem*. The objective of citation cycling is to build a coherent picture or overview of the literature relating to some topic [3]. Current hypertext interfaces employ a simple paging model, in which the user navigates the citation graph in a step-wise fashion from one node of interest to another. Typically the user begins by viewing an article summary before scrolling down or clicking through to a list of cited or citing articles, from which they can then select and navigate through to a new article summary, and so on. As they move from one node to the next they will tend to follow many interweaving paths within the citation graph. Yet the user's view is always narrowly focused on either the current article, its citations or citing articles. As such, the user's awareness of the developing context – the relative salience of relevant articles and their inter-relationships – is dependent upon their ability to assimilate these many local views into a coherent a mental model of the space.

In this paper, we propose a new interaction model for citation cycling which addresses these and other limitations of associated with current citation index interfaces. The model employs a three-list view to represent the aggregated sets of first-order citations (one generation back and forth) associated with the set of one or more known relevant articles. Known, relevant article records are displayed in the central 'pearl' list, whilst the left- and the right-hand lists display all unique articles that are cited by and cite the pearl articles. Following the cognitive overlap hypothesis [5, 6], articles that relate to more known articles are deemed more salient. As more articles are added to the pearl list, clear differences emerge in terms of the number of relational 'hits' they make with the pearl, providing valuable relevance feedback to the user. We call this interaction model Citation Chain Aggregation (CCA).

2. RELATED WORK

Given the focus and context problem, the idea of visualizing relevant sub-graphs of the citation index has been explored by several studies (e.g. Butterfly: [7]; Circleview: [2]; CAVis: [9] and [11]). Perhaps the earliest and best known example is the Butterfly visualizer [7]. The body of each butterfly object is used to display metadata about a specific article record, whilst the wings display annotated links to cited (left) and citing (right) articles. Clicking on a link creates a new butterfly instance focusing on that article. Citation chains are conveyed by folding-away the wings and placing butterfly objects side by side, again with the cited articles to the left of citing items.

The idea of representing a single generation citation chain in *cited-focus-citing* order (see also [11]) is a natural mapping that has been translated into the CCA prototype design. However, a pervasive issue associated with all approaches to visualizing citation graphs is one of controlling visual clutter. The local citation graph surrounding just a few related articles can be large and complex. Representing such a graph neatly and usably in iconic format is problematic enough, but effective navigational support also requires details in context. Finding the balance between overview and sufficient detail typically restricts the size of the visible context to just one current focus article and one or two generations of ancestors and descendants [2, 11]. Also, when the cardinality is high, less salient citation links must be either hidden (e.g. [2, 7]) or suppressed in some way [e.g. 11].

We contend that a solution to the clutter problem may be to forgo the convention of visualizing the graph *per se* and to focus instead on a more abstract representation of the evolving context. The interaction model proposed here draws inspiration from work in the area of information retrieval, specifically Ingwersen's principle of polyrepresentation [5] and related work by Larsen [6]. Studies by McCain [8] and Pao [10] have shown that the overlap in results sets retrieved from multiple, cognitively distinct queries tend to identify the most relevant articles. For instance, Pao [10] found that the overlap between documents retrieved from both a keyword search on MEDLINE and a citation search increased precision by some six to eight times.

More recently, Larsen [6] demonstrated how the "boomerang" effect can improve search precision, by combining keyword and citation search without the need to provide expert-defined seed articles. Larsen proposed a three step method. In step one, multiple queries are made to separate indexes (e.g. title, keywords and abstract). Each search will tend to produce different results. However the *cognitive overlap hypothesis* predicts that citations common to multiple sets will tend to be the most important. As such, in step two, only citations made by articles in two or more of the step one sets are retained. In step three, a forward citation search is performed for all of the retained citations. Finally, these citing articles are ranked by the number of step two sub-sets they relate to. Using this method, Larsen found that rated precision tended to increase significantly inline with the degree of overlap.

3. CITATION CHAIN AGGREGATION

Larsen's method represents an effective proxy for citation cycling. However, CCA improves upon this approach by being more interactive and open-ended. Larsen's aim was to reduce inconsistency resulting from the need to select appropriate 'seed' documents [8]. In contrast, we see the task of 'pearl growing' from a one or two seed articles as a natural and likely scenario [1, 4]. As such, in the CCA model, the user has complete control over which articles participate in the cycling process. For example, the user can begin with a single article, select just a handful of key citations, simultaneously forward chain from all of these articles and immediately see the resulting overlaps in terms of both citing and cited articles. The user is then free to select whichever articles they wish, from either side, to continue the cycling process. Note that the model is also flexible enough to emulate the boomerang strategy in some form, although this isn't currently implemented in the prototype.

The fundamental concept in CCA is the *citation chain*. We define a citation chain as the sub-graph describing an article and all of its immediate ancestors (cited articles) and descendents (citing articles). In the conventional hypertext model, the user can only view one citation chain (or part-thereof) at a time. CCA, on the other hand, aggregates multiple citation chains into a single three-list view (Figure 1).

A citation cycling episode begins with the researcher adding one or more relevant articles to the pearl list. This can be achieved by entering a unique identifier or by importing results from a search (or any other list of references). When an article is added, the system sends a query to the citation index web service to retrieve all its cited and citing articles. The results are then

displayed in the requisite peripheral lists. If an article is not already in its target list then a new item is added, otherwise the item's counter is incremented to reflect another instance of overlap or 'hit' between citation chains. Note, therefore, that any article can be a member of any number of lists.

Figure 1 shows the result after just two related articles have been added to the pearl list. The seed article was added first (upper black circle), followed by one of its citations, selected from the resulting items in the cited list. The Venn diagram shows that many of the citations, both to and from the two pearl articles, are shared. The sketch below shows how the results are displayed in the list views. Each list item comprises various fields. In addition to bibliographic data (author, date etc) the left hand field in each view shows an overlap counter. We can see that half of the articles in the citing list cite both pearl articles (count of two), whilst half of the cited items are cited by both articles. Following the cognitive overlap hypothesis, list items can be sorted from high to low by this 'hit' counter to provide relevance feedback. As more items are added to the pearl list, so the differentiation between unseen articles increases. Within the pearl list, an item's counter field represents the number of pearl items that cite it. In this case, the second article is attributed a count of one. Gradually as more articles are added, the differentiation that emerges here provides a measure of salience within the pearl list.

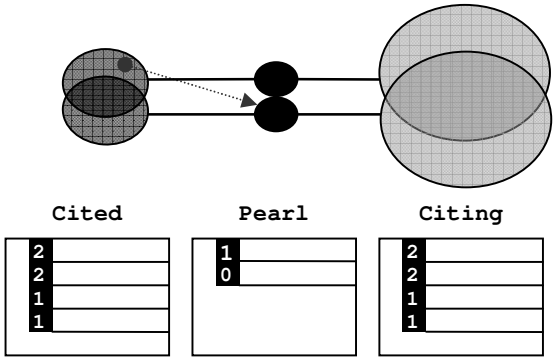

Figure 1. Citation chain aggregation of two articles.

4. PROTOTYPE

The prototype required direct and reliable access to a comprehensive online citation index. The first prototype utilized the open REST API of CiteSeerX (http://citeseerx.ist.psu.edu). The open nature of the API was convenient for an initial proof-of-concept, but the database lacked the coverage required for field testing, being mainly focused on the CS literature. Permission was therefore obtained from Thomson-Reuters to use their WOS SOAP API, which enabled a more widely useful prototype to be developed. Both versions have almost identical interfaces and functionality, sharing much of the same code in terms of presentation and interaction routines. The key differences lie in the implementation of search requests and handling of the responses.

The screenshot shown in Figure 2 is from the WOS version. The three lists are situated at the top of the interface. By default, lists are sorted from high to low by the overlap counter field (first column). Lists can be re-sorted by any field by clicking on the column header. The user can click on any list item to view the article record, which is displayed underneath in a text box. Where available, the user can link directly to the publisher's web-site to view the full text article. Tight coupling is employed to convey relationships. Clicking on a cited/citing item causes related records in the pearl list to be highlighted in bold. Likewise, clicking on a pearl item highlights related records in the peripheral lists. Occurrences of pearl members in the peripheral lists are de-emphasized in grey. Seed articles can be located by executing a WOS search. Search results can be viewed and directly transferred to the pearl list from within the interface. Currently topic, title, author and date search fields are supported.

Internally, all article records are stored within a single array defined by a custom type. This type is composed of fields describing the article's bibliographic data and set membership (cited, pearl and citing). Note that a record can belong to any number of sets. Cited and citing status fields are, in turn, defined by a sub-type that maintains a count and index of any relationships with pearl articles.

Server response times are the only bottleneck when it comes to performance, particularly when requesting citing articles, which must be retrieved in batches of 100 records. However, most updates complete within just a few seconds. Updates that do not involve server requests remain near-instant even when the array contains thousands of items.

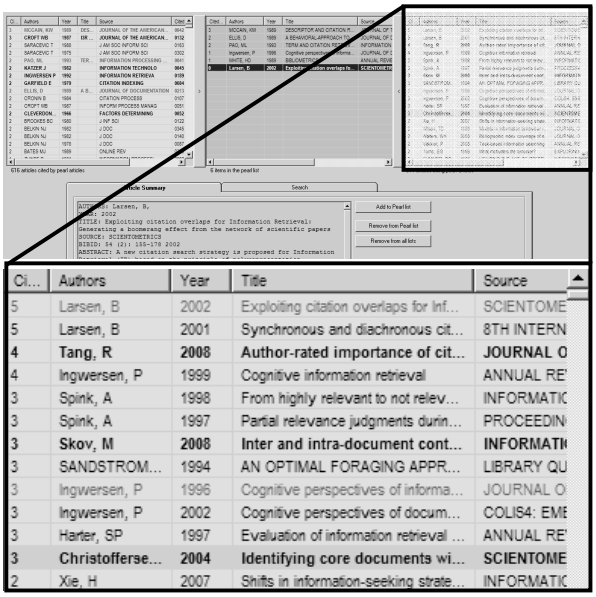

Figure 2. Prototype system interface (top) with close-up view of citing articles list (bottom).

5. WALKTHROUGH

We have yet to complete the first round of formal user testing. However, the utility of CCA can be demonstrated informally by walking through a single citation cycling iteration.

A convenient example is to use Larsen's [6] paper on the 'boomerang effect' as a seed. This cites 31 articles in total, but we begin by selecting just a handful of citations that seem most pertinent to our interests. First Ellis [4], as the seminal paper

about citation chaining, is transferred to the pearl. Next, Ingwersen's [5] paper on polyrepresentation, which evidently forms the theoretical basis for Larsen's retrieval model, is added. We back this up with McCain [8] and Pao [10] which provide empirical evidence in support of the cognitive overlap hypothesis. Finally, we identify a fifth citation, White and McCain [12], a seminal paper on co-citation analysis, in the hope of picking up other work that has investigated the relationship between co-citation and search.

The result is a total of 611 cited articles and 616 citing articles. Clearly too many to browse through exhaustively! However, the sorted citing list reveals that only a handful (<10%) of items cite more than one pearl article. In fact just 12 (2%) cite three or more of the six pearl articles (see Table 1), with a further 47 (7.7%) citing two.

WOS related records (top-ten only) and citing articles (total of six) searches were requested for the seed article. Table 1 compares the results with the CCA rankings. The top CCA rank is, predictably, the seed itself. The next article, also with an overlap score of five, is a conference paper on the boomerang effect co-authored by Larsen and Ingwersen, which predates the seed but was not cited by it. This is followed by Tang and then Ingwersen, each with scores of four. Tang is particularly interesting as it was ranked lowest amongst the related records (10th) and was only the third highest cited (two cites) amongst those citing the seed.

Citing List	Overlap	Related to seed	Cites Seed
Larsen (2002) Sciento.	5		
Larsen (2001) ISSI-2001	5	✓	
Tang (2008) J. Doc	4	✓	✓
Ingwersen (1999) ARIST	4	✓	
Spink (1998) IP&M	3		
Spink (1997) Proc. ASIS	3		
Skov (2008) IP&M	3		✓
Sandstrom (1994) Lib. Quart.	3		
Ingwersen (1996) J. Doc	3	✓	
Ingwersen (2002) COLIS 4	3		
Harter (1997) ARIST	3	✓	
Christoffersen (2004) Sciento.	3	✓	✓

Table 1: Articles citing three or more pearl articles

More generally, it is interesting to see how some items are promoted whilst others are demoted. Six out of the top-ten related items remain highly ranked after CCA. The dropped items were all science mapping, rather than information retrieval studies, and so were indeed less relevant. Finally, a further four articles are new entries that were not picked up by either of the comparison searches.

Bear in mind that this example is the result after just a single iteration. The CCA model allows for an open-ended interaction with the citation index. The user can therefore engage in a game of 'citation tennis', cycling backwards and forwards, until they are satisfied with the contents of their pearl list.

6. CONCLUSIONS AND FUTURE WORK

This paper has proposed a new interaction model to support citation cycling. CCA resolves key issues associated with existing hypertext- and visualization-based interfaces. The most notable advantages are its simplicity and scalability, both in terms of system requirements and the user experience. We have demonstrated informally how just a single backward-forward chaining cycle can provides clear and useful relevance feedback on unseen articles. This analysis needs to be followed-up with formal empirical studies. Work is in progress to evaluate the benefits to researchers as they use the system both in the laboratory and over extended periods in the field.

Future work will also explore ways of improving the prototype with additional functionality. There are many possible enhancements that can be made to support interaction within larger citation graphs, including content similarity ranking, visualization and dynamic filtering.

Finally, CCA is a flexible model that may be adapted and optimized for a variety of different information tasks, ranging from exploration through to checking the completeness of referencing during editing and review. Moreover, its use is not limited only to academic literature search, but potentially any information space that can be represented as a directed graph. Obvious domains for future evaluation include legal search, patent analysis (prior-art search) and web search.

7. REFERENCES

[1] Bates, M. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online Review,* 13, 5, 407-431.

[2] Bergström, P. and Whitehead, E.J. 2006. CircleView: scalable visualization and navigation of citation networks. In *Proceedings of 2006 Symposium on Interactive Visual Information Collections and Activity* (College Station, TX).

[3] Cawkell, A.E. 1998. Checking research progress on 'image retrieval by shape-matching' using the Web of Science. *Aslib Proceedings*, 50, 2, 27-31.

[4] Ellis, D. 1989. A behavioural approach to information retrieval system design. *Journal of Documentation*, 45, 3, 171-212.

[5] Ingwersen, P. 1996. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, 52, 1, 3-50.

[6] Larsen, B. 2002. Exploiting citation overlaps for information retrieval: generating a boomerang effect from the network of scientific papers. *Scientometrics*, 54, 2. 155-178.

[7] Mackinlay, J.D., Rao, R. and Card, S.K. 1995. An organic user interface for searching citation links. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems* (Denver, CO). 67-73.

[8] McCain, K.W. 1989. Descriptor and citation retrieval in the medical behavioral sciences literature: Retrieval overlaps and novelty distribution. *Journal of the American Society for Information Science*, 40, 2, 110-114.

[9] Nguyen, Q.V., Huang, M.L. and Simo, S. 2007. Visualization of relational structure among scientific articles.

In *Proceedings of the 9th International Conference on Visual Information Systems* (Shanghai, China). 415-425.

[10] Pao, M.L. 1993. Term and citation retrieval - a field-study. *Information Processing & Management*, 29, 1. 95-112.

[11] Schäfer, U. and Kasterka, U. 2010. Scientific authoring support: a tool to navigate in typed citation graphs. In *Proceedings of NAACL HLT 2010 Workshop on Computational Linguistics and Writing* (Los Angeles, CA). 7-14.

[12] White, H. and McCain, K. 1998. Visualizing a discipline: an author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, 49, 4. 327-355.