Rapid Understanding of Scientific Paper Collections: Integrating Statistics, Text Analysis, and Visualization

Cody Dunne[1,3*], Ben Shneiderman[1,3], Robert Gove[1,3], Judith Klavans[2] and Bonnie Dorr[2,3]

[1]Human-Computer Interaction Lab, [2]Computational Linguistics and Information Processing Lab [3]Department of Computer Science
University of Maryland, College Park, MD 20742.
E-mail: {cdunne, ben, rpgove, bonnie}@cs.umd.edu, jklavans@umd.edu
[*]Corresponding author

Abstract

Keeping up with rapidly growing research fields, especially when there are multiple interdisciplinary sources, requires substantial effort for researchers, program managers, or venture capital investors. Current theories and tools are directed at finding a paper or website, not gaining an understanding of the key papers, authors, controversies, and hypotheses. This report presents an effort to integrate statistics, text analysis, and visualization in a multiple coordinated window environment that supports exploration. Our prototype system, Action Science Explorer (ASE), provides an environment for demonstrating principles of coordination and conducting iterative usability tests of them with interested and knowledgeable users. We developed an understanding of the value of reference management, statistics, citation context extraction, natural language summarization for single and multiple documents, filters to interactively select key papers, and network visualization to see citation patterns and identify clusters. The three-phase usability study guided our revisions to ASE and led us to improve the testing methods.

Introduction

Contemporary scholars and scientists devote substantial effort to keep up with advances in their rapidly expanding fields. The growing number of publications combined with increasingly cross-disciplinary sources makes it challenging to follow emerging research

The authors would like to thank Dragomir Radev and Vahed Qazvinian for data preparation and Jimmy Lin for discussion of this work. This work is funded by the National Science Foundation grant "iOPENER: A Flexible Framework to Support Rapid Learning in Unfamiliar Research Domains" jointly awarded to the University of Michigan and University of Maryland as IIS 0705832.

fronts and identify key papers. It is even harder to begin exploring a new field without a starting frame of reference.

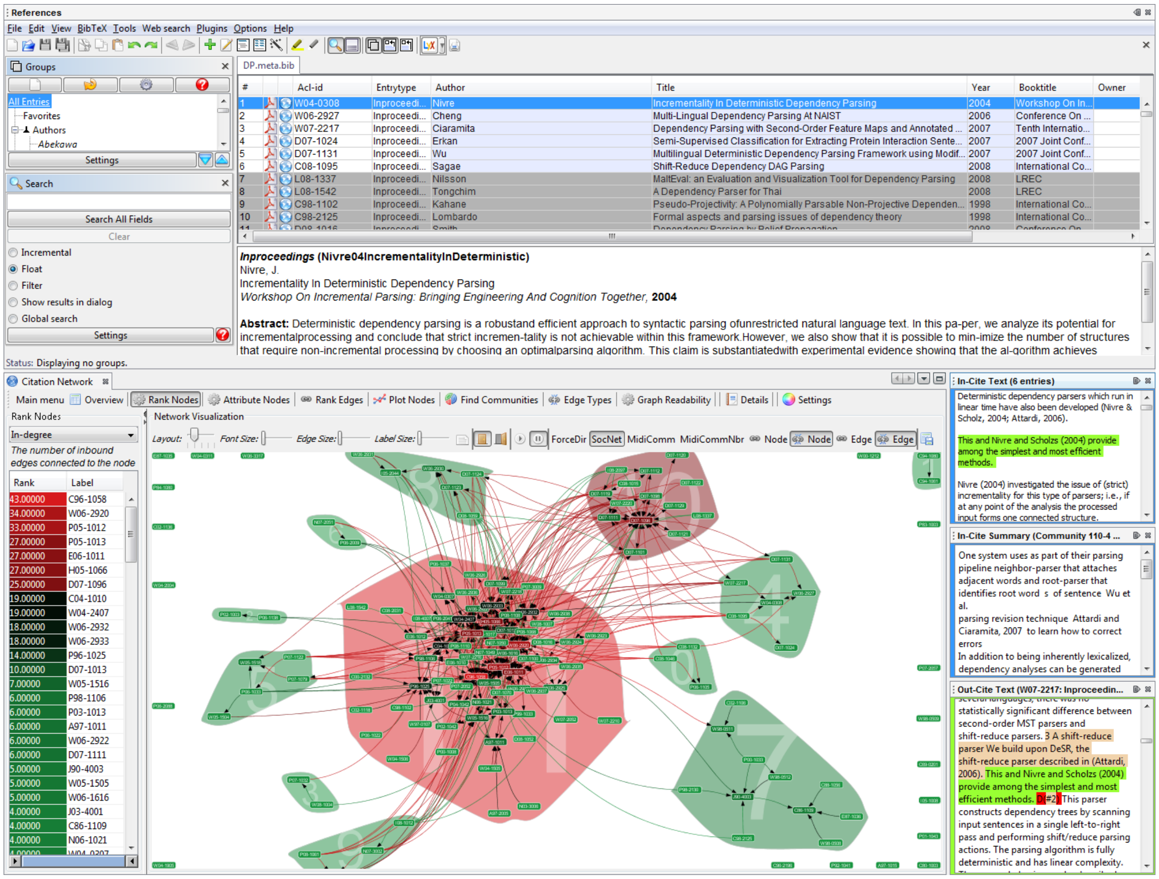
Researchers have vastly different levels of expertise and requirements for learning about scientific fields. A graduate student or cross-disciplinary researcher in a new field might find it useful to see the pivotal historical papers, key authors, and popular publication venues. On the other hand, a seasoned academic may be interested only in recent leading work and outlier papers or authors that challenge their preconceptions about the field. Grant program managers and review panel members sometimes have to examine fields they are not familiar with, looking for research trends, emerging fields, and open questions. Moreover, social scientists or scientometric analysts may be interested in how academic communities form over time, comparing citation and publication trends by country, or tracking the adoption of a single innovation.

Tools for rapid exploration of the literature can help ease these difficulties, providing readers with concise overviews tailored to their needs and aiding the generation of accurate surveys. Digital libraries and search engines are useful for finding particular papers or those matching a search string, but do not provide the additional analysis tools required to quickly summarize a field. Users unfamiliar with the field often find it challenging to search out the influential or groundbreaking papers, authors, and journals.

Specialized tools compute statistical measures and rankings to help identify items of interest, and other tools automatically summarize the text of multiple papers to extract key points. However, all these tools are decoupled from the literature exploration task and are not easily integrated into the search process. Visualization techniques can be used to provide immediate overviews of publication and citation patterns in a field, but are uncommon in literature exploration tools. When present, they usually do not display enough data or provide the interaction techniques required to analyze the publication trends and research communities in a field. Even more ambitiously the goals include sufficient understanding to enable decision-making such as which fields are promising directions for researchers, appropriate for increased/reduced funding by government or industrial program managers, or worthy of investment by a venture capital organization.

This paper presents the results of an effort to integrate statistics, text analysis and visualization in a powerful prototype interface for researchers and analysts. The Action Science Explorer[1] (ASE) is designed to support exploration of a *collection* of *papers* so as to rapidly provide a summary, while identifying key papers, topics, and research groups. ASE uses 1) bibliometric lexical link mining to create a citation network for a field and context for each citation, 2) automatic summarization techniques to extract key points from papers, and 3) potent network analysis and visualization tools to aid in the exploration relationships. ASE, shown in Fig. 1, presents the academic literature for a field using many different modalities: tables of papers, full texts, text summaries, and visualizations of the structure of the citation network and the groups it contains. Each view of the underlying data is coordinated such that papers selected in one view are highlighted in the others, providing additional metadata, text summaries, and statistical measure rankings about them. Users can filter by rankings or via search queries, highlighting the matching results in all views.

[1] For videos and more information visit http://www.cs.umd.edu/hcil/ase



#	Acl-id	Entrytype	Author	Title	Year	Booktitle	Owner
1	W04-0308	Inproceedi...	Nivre	Incrementality In Deterministic Dependency Parsing	2004	Workshop On In...	
2	W06-2927	Inproceedi...	Cheng	Multi-Lingual Dependency Parsing At NAIST	2006	Conference On ...	
3	W07-2217	Inproceedi...	Ciaramita	Dependency Parsing with Second-Order Feature Maps and Annotated ...	2007	Tenth Internatio...	
4	D07-1024	Inproceedi...	Erkan	Semi-Supervised Classification for Extracting Protein Interaction Sente...	2007	2007 Joint Conf...	
5	D07-1131	Inproceedi...	Wu	Multilingual Deterministic Dependency Parsing Framework using Modif...	2007	2007 Joint Conf...	
6	C08-1095	Inproceedi...	Sagae	Shift-Reduce Dependency DAG Parsing	2008	International Co...	
7	L08-1337	Inproceedi...	Nilsson	MaltEval: an Evaluation and Visualization Tool for Dependency Parsing	2008	LREC	
8	L08-1542	Inproceedi...	Tongchim	A Dependency Parser for Thai	2008	LREC	
9	C98-1102	Inproceedi...	Kahane	Pseudo-Projectivity: A Polynomially Parsable Non-Projective Dependen...	1998	International Co...	
10	C98-2125	Inproceedi...	Lombardo	Formal aspects and parsing issues of dependency theory	1998	International Co...	
11	D08-1016	Inproceedi...	Smith	Dependency Parsing by Belief Propagation	2008	Conference On...	

***Inproceedings* (Nivre04IncrementalityInDeterministic)**
Nivre, J.
Incrementality In Deterministic Dependency Parsing
Workshop On Incremental Parsing: Bringing Engineering And Cognition Together, **2004**

Abstract: Deterministic dependency parsing is a robustand efficient approach to syntactic parsing ofunrestricted natural language text. In this pa-per, we analyze its potential for incrementalprocessing and conclude that strict incremen-tality is not achievable within this framework.However, we also show that it is possible to min-imize the number of structures that require non-incremental processing by choosing an optimalparsing algorithm. This claim is substantiatedwith experimental evidence showing that the al-gorithm achieves

Status: Displaying no groups.

Rank Nodes
In-degree
The number of inbound edges connected to the node

Rank	Label
43.00000	C96-1058
34.00000	W06-2920
33.00000	P05-1012
27.00000	P05-1013
27.00000	E06-1011
27.00000	H05-1066
25.00000	D07-1096
19.00000	C04-1010
19.00000	W04-2407
18.00000	W06-2932
18.00000	W06-2933
14.00000	P96-1025
10.00000	D07-1013
7.00000	W05-1516
6.00000	P98-1106
6.00000	P03-1013
6.00000	A97-1011
6.00000	W06-2922
6.00000	D07-1111
5.00000	J90-4003
5.00000	W05-1505
5.00000	W06-1616
4.00000	J03-4001
4.00000	C86-1109
4.00000	N06-1021
4.00000	W04-0307

In-Cite Text (6 entries)

Deterministic dependency parsers which run in linear time have also been developed (Nivre & Scholz, 2004; Attardi, 2006).

This and Nivre and Scholzs (2004) provide among the simplest and most efficient methods.

Nivre (2004) investigated the issue of (strict) incrementality for this type of parsers; i.e., if at any point of the analysis the processed input forms one connected structure.

In-Cite Summary (Community 110-4 ...

One system uses as part of their parsing pipeline neighbor-parser that attaches adjacent words and root-parser that identifies root word s of sentence Wu et al.
parsing revision technique Attardi and Ciaramita, 2007 to learn how to correct errors
In addition to being inherently lexicalized, dependency analyses can be generated

Out-Cite Text (W07-2217: Inproceedin...

several languages, there was no statistically significant difference between second-order MST parsers and shift-reduce parsers. 3 A shift-reduce parser We build upon DeSR, the shift-reduce parser described in (Attardi, 2006). This and Nivre and Scholzs (2004) provide among the simplest and most efficient methods. D(#2) This parser constructs dependency trees by scanning input sentences in a single left-to-right pass and performing shift/reduce parsing actions. The parsing algorithm is fully deterministic and has linear complexity.

Figure 1. : The main interface of Action Science Explorer

Related Work

To accomplish the goals laid out in the introduction, a complete system needs to support a variety of services. Initially users would search a large collection and import the relevant papers to deepen their understanding of the desired scientific field. Most research database systems support searching the collection and return a list of papers, but only a few provide sufficiently powerful tools to explore the result set. Natural operations would be to sort and filter the result set by time, author name, institutions, key phrases, search term relevance, citation frequency, or other impact measures. These help users to identify the key papers, researchers, themes, research methods, and disciplinary links as defined by publication venue.

As users invest time to gain familiarity with individual papers, they study the list of authors, read the abstracts, scan the content, and review the list of citations to find familiar papers, authors, and journals. Another source of insight about a paper is to see how later papers describe it and to see what other papers are cited concurrently. Studying such citation context is a fruitful endeavor, but is difficult in most systems.

After studying 5–50 papers users usually begin to understand the field, key researchers, consistent topics, controversies, and novel hypotheses. They may annotate the papers, but more commonly they put them into groups to organize their discovery process and facilitate future usage. Accelerating the process of gaining familiarity would yield enormous benefits, but a truly helpful system would also improve the completeness, appropriateness, and value of the outcome.

Once users have gained familiarity they may dig deeper to understand the major breakthroughs and remaining problems. Breakthroughs and problems are rarely spelled out explicitly as a field is emerging, although review papers that look back over a decade or two are likely to contain such insights. Reading citation contexts is helpful for gaining insights into the field, but can be time consuming even in a well-designed system and might give only a narrow focus. Ranked lists, standard charts, and scatterplots can provide useful overviews, but citation network visualizations can potentially generate a higher payoff. Network visualizations have been only marginally effective in the past, but improved layout, clustering, ranking, statistics, and filtering techniques have the potential for exposing patterns, clusters, relationships, gaps, and anomalies. Even more potent for those studying emerging fields is the capacity to explore an evolutionary visualization that shows the appearance of an initial paper, the gradual increase in papers that cite it, and sometimes the explosion of activity for "hot" topics. Other temporal phenomena are the cross disciplinary citations, fracturing of research topics, and sometimes the demise of a hypotheses.

Techniques of natural language processing can potentially speed up the analysis of a large collection by extracting frequently occurring terms/phrases, identifying topics, and identifying key concepts. Multi-document summarization and document clustering have the potential to help users by providing some forms of automated descriptions for interesting subsets of a collection.

Since accomplishing these complex tasks in a single scrolling window is difficult, many systems provide multiple coordinated windows that enable users to see lists or visualizations in one window and make selections for displays in other windows. A more advanced technique is brushing and linking, which allows selection in one display to highlight related items in another display.

Existing systems provide some of these features in various combinations, though none allow users to leverage all of them in a single analysis. For their initial exploration, users frequently use academic search tools like Google Scholar (Google, 2011) and Microsoft Academic Search (Microsoft Research, 2011). Subscriber-only general databases are used frequently at universities and research labs, such as ISI Web of Knowledge (Thomson Reuters, 2011b) and SciVerse Scopus (Elsevier, 2011). Additionally, many field-specific databases exist such as PubMed (National Center for Biotechnology Information, 2011) for Life and Biological Sciences. Computer and Information Sciences have databases like the web harvesting CiteSeer (Giles, Bollacker, & Lawrence, 1998; Bollacker, Lawrence, & Giles, 1998), arXiv (Cornell University Library, 2011) for preprints, and the publisher-run ACM Digital Library (Association for Computing Machinery, 2011) and IEEE Xplore (Institute of Electrical and Electronics Engineers, 2011).

These search tools and databases generally provide a sortable, filterable list of papers matching a user-specified query, sometimes augmented by faceted browsing capabilities and general overview statistics. Some enable users to save specific papers into groups to review or export later, though this is via a separate interface and annotation is not usually supported. ISI Web of Knowledge is rare in that it includes a visualization of the ego network of an individual paper, including both incoming and outgoing citations. However, it is a hyperbolic tree visualization that has little dynamic interaction. Furthermore, visualizations are most useful for finding overall trends, clusters, and outliers–not for looking at small ego network subsets.

An emerging category of products called reference managers enhances these paper management capabilities by supporting additional search, grouping, and annotation features, as well as basic collection statistics or overview visualizations. Some examples are JabRef (JabRef Development Team, 2011), Zotero (Center for History and New Media, 2011), EndNote (Thomson Reuters, 2011a), and Mendeley (Mendeley Ltd. 2011).

Many academic databases now use citation extraction to help build the citation network of their paper collections for bibliometric analysis, and some such as CiteSeer (Giles, 1998; Bollacker, 1998) and Microsoft Academic Search (Microsoft Research, 2011) expose the context of those citations. The benefit of showing citation context is that readers can quickly learn about the critical reception, subsequent and similar work, and key contributions of a paper as seen by researchers later on. Analyses of paper collections from citation text has also been demonstrated to be useful for a wide range of applications. Bradshaw (2003) used citation texts to determine the content of papers and improve the results of a search engine. Even the author's reason for citing a given paper can be automatically determined (Teufel, Siddharthan, & Tidhar, 2006).

Natural language processing techniques for document and multi-document summarization can produce distilled output that is intended to capture the deeper meaning behind a topically grouped set of papers. Citation texts have been used to create summaries of single papers (Qazvinian & Radev, 2008; Mei & Zhai, 2008). Nanba and Okumura (1999) discuss citation categorization to support a system for writing surveys and Nanba, Abekawa, Okumura, and Saito (2004) automatically categorize citation sentences into three groups using pre-defined phrase-based rules. Other summarization approaches exist for papers (Teufel & Moens, 2002) or news topics (Radev, Otterbacher, Winkel, & Blair-Goldensohn, 2005). For a cogent review of summarization techniques, see Sekine and Nobata (2003).

Academic research tools apply bibliometrics to help users understand collections through network visualizations of paper citations, author collaborations, author or paper co-citations, and user access patterns. (Newman, 2001) Many standard bibliometric analysis and visualization approaches are integrated in Network Workbench (NWB Team, 2006). Another tool designed for analyzing evolving fields is CiteSpace (Chen, 2004; Chen, 2006; Chen, Ibekwe-SanJuan, & Hou, 2010), which is targeted at identifying clusters and intellectual turning points. Similarly, semantic substrates can be used for citation network visualization (Aris, Shneiderman, Qazvinian, & Radev, 2009), showing scatterplot layouts of nodes to see influence between research fronts. Unfortunately these visualizations are weakly integrated into the rest of the exploration process and are yet to be widely used.

Part of the challenge of integrating visualizations effectively is making them visible concurrently with the search result list. Effective designs would move from the traditional single scrolling windows to *multiple coordinated views* that support brushing and linking to highlight related items (North & Shneiderman, 1997; Shneiderman, Plaisant, Cohen, & Jacobs, 2009). The power of a spatially stable overview and multiple detail views is especially appropriate for browsing large collections of papers. However, many bibliometrics tools that present several views of the collection would benefit from better integration, easier linking, and common user interfaces across windows (e.g., Network Workbench (NWB Team, 2006)).

Existing theories of information seeking are helpful for reminding us of process models that start from identifying the goal and end with presenting the results to others (Hearst, 2009). One example is Kuhlthau's six stages: initiation, selection, exploration, formulation, collection, and presentation (Kuhlthau, 1991). Marchionini (1997) describes an 8-stage process in his early book, and offers a richer model in his more recent descriptions of exploratory search (Marchionini, 2006). These and other information-seeking processes (Bates, 1990) provide a useful foundation for the complex task of enabling users to understand emerging fields. This complex task also benefits from theories of sense-making and situation awareness, since the goal is to understand multiple aspects of emergent fields such as the key papers, authors, controversies, and hypotheses. A related goal is to understand the relation to other fields which could be sources of insight and fields which have parallel or duplicate results that are not recognized. A further goal is to determine which topics have the greatest potential for advancing a field, thereby guiding researchers, program managers at funding agencies, or venture capital investors who see commercial potential.

ASE Design

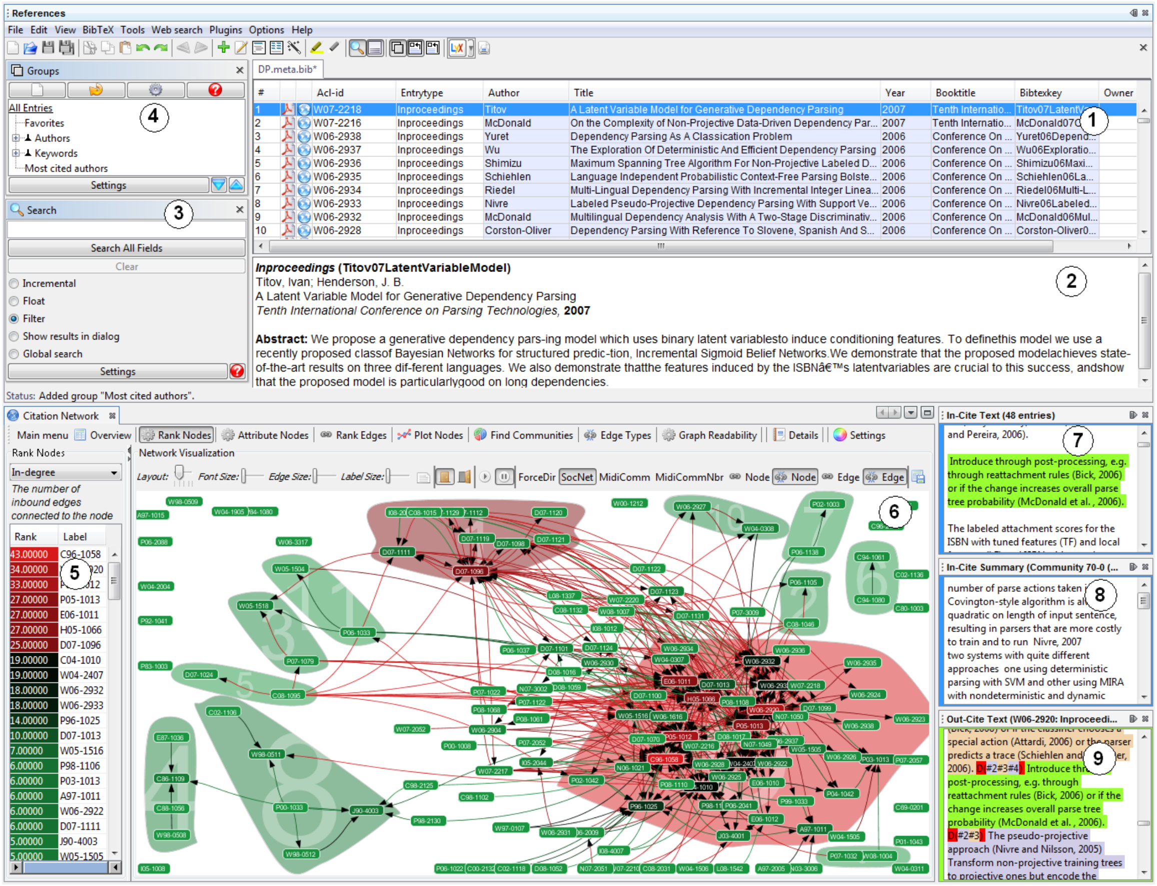
The goal of Action Science Explorer (ASE) is to help analysts rapidly generate readily-consumable surveys, of emerging research topics or fields they are unfamiliar with, targeted to different audiences and levels. The philosophy of our design is to integrate statistical, visual, and textual representations that are each relevant to the task of scientific literature exploration. All of these modalities are linked together in multiple coordinated views, such that any selection in one is reflected in the others. This section describes the design and various features of ASE, illustrated in Fig. 2.

Search & Data Import

We build on familiar literature exploration interfaces: the search engines and databases often used when conducting literature reviews. A typical ASE session begins with a keyword, phrase, or topic search of a database to define a target corpus that is retrieved and processed. In our examples, we use the 147 papers returned by a search for "Dependency Parsing" on the ACL Anthology Network (AAN) (Radev, Muthukrishnan, & Qazvinian, 2009; Radev, Joseph, Gibson, & Muthukrishnan, 2009), a collection of 16,857 Computational Linguistics papers. The collection includes the full text of each paper from OCR extraction and manual cleanup, from which the authors and citations between papers had been retrieved automatically. The authors and citations had required substantial cleaning, disambiguation, and correction which were done manually, assisted by an n-best matching algorithm with $n = 5$.

Reference Management

The search results are loaded into ASE and displayed using the JabRef reference manager (JabRef Development Team, 2011) component, shown in Fig. 2 (1–4). This provides users with a table of papers and their bibliographic data (1), from which the URL or DOI, full-text PDF, plain text, and any other files for each paper can be opened. The citation version of the selected paper is shown along with its abstract and any user-written



#	Acl-id	Entrytype	Author	Title	Year	Booktitle	Bibtexkey	Owner
1	W07-2218	Inproceedings	Titov	A Latent Variable Model for Generative Dependency Parsing	2007	Tenth Internatio...	Titov07LatentV...	
2	W07-2216	Inproceedings	McDonald	On the Complexity of Non-Projective Data-Driven Dependency Par...	2007	Tenth Internatio...	McDonald07C...	
3	W06-2938	Inproceedings	Yuret	Dependency Parsing As A Classication Problem	2006	Conference On ...	Yuret06Depend...	
4	W06-2937	Inproceedings	Wu	The Exploration Of Deterministic And Efficient Dependency Parsing	2006	Conference On ...	Wu06Exploratio...	
5	W06-2936	Inproceedings	Shimizu	Maximum Spanning Tree Algorithm For Non-Projective Labeled D...	2006	Conference On ...	Shimizu06Maxi...	
6	W06-2935	Inproceedings	Schiehlen	Language Independent Probabilistic Context-Free Parsing Bolste...	2006	Conference On ...	Schiehlen06La...	
7	W06-2934	Inproceedings	Riedel	Multi-Lingual Dependency Parsing With Incremental Integer Linea...	2006	Conference On ...	Riedel06Multi-L...	
8	W06-2933	Inproceedings	Nivre	Labeled Pseudo-Projective Dependency Parsing With Support Ve...	2006	Conference On ...	Nivre06Labeled...	
9	W06-2932	Inproceedings	McDonald	Multilingual Dependency Analysis With A Two-Stage Discriminativ...	2006	Conference On ...	McDonald06Mul...	
10	W06-2928	Inproceedings	Corston-Oliver	Dependency Parsing With Reference To Slovene, Spanish And S...	2006	Conference On ...	Corston-Oliver0...	

Inproceedings **(Titov07LatentVariableModel)**
Titov, Ivan; Henderson, J. B.
A Latent Variable Model for Generative Dependency Parsing
Tenth International Conference on Parsing Technologies, **2007**

Abstract: We propose a generative dependency pars-ing model which uses binary latent variablesto induce conditioning features. To definethis model we use a recently proposed classof Bayesian Networks for structured predic-tion, Incremental Sigmoid Belief Networks.We demonstrate that the proposed modelachieves state-of-the-art results on three dif-ferent languages. We also demonstrate thatthe features induced by the ISBNâ€™s latentvariables are crucial to this success, andshow that the proposed model is particularlygood on long dependencies.

Status: Added group "Most cited authors".

Rank Nodes: In-degree
The number of inbound edges connected to the node

Rank	Label
43.00000	C96-1058
34.00000	920
33.00000	P 012
27.00000	P05-1013
27.00000	E06-1011
27.00000	H05-1066
25.00000	D07-1096
19.00000	C04-1010
19.00000	W04-2407
18.00000	W06-2932
18.00000	W06-2933
14.00000	P96-1025
10.00000	D07-1013
7.00000	W05-1516
6.00000	P98-1106
6.00000	P03-1013
6.00000	A97-1011
6.00000	W06-2922
6.00000	D07-1111
5.00000	J90-4003
5.00000	W05-1505

In-Cite Text (48 entries)

and Pereira, 2006).

Introduce through post-processing, e.g. through reattachment rules (Bick, 2006) or if the change increases overall parse tree probability (McDonald et al. , 2006).

The labeled attachment scores for the ISBN with tuned features (TF) and local

In-Cite Summary (Community 70-0 (...

number of parse actions taken i Covington-style algorithm is al quadratic on length of input sentence, resulting in parsers that are more costly to train and to run Nivre, 2007
two systems with quite different approaches one using deterministic parsing with SVM and other using MIRA with nondeterministic and dynamic

Out-Cite Text (W06-2920: Inproceedi...

special action (Attardi, 2006) or the parser predicts a trace (Schiehlen and er, 2006). D(#2#3#4) Introduce thr post-processing, e.g. through reattachment rules (Bick, 2006) or if the change increases overall parse tree probability (McDonald et al. , 2006). D(#2#3) The pseudo-projective approach (Nivre and Nilsson, 2005) Transform non-projective training trees to projective ones but encode the

Figure 2. : The main views of ASE are displayed and labeled here: Reference Management (1–4), Citation Network Statistics & Visualization (5–6), Citation Context (7), Multi-Document Summaries (8), and Full Text with hyperlinked citations.

annotations (2), and additional metadata can be shown or entered by double-clicking on an entry. The table can be sorted by column and searched using regular expressions (3), and papers can be organized into hierarchical overlapping groups.

As the underlying data structure used by JabRef is the BibTeX bibliography format, ASE can be easily used in conjunction with LaTeX and, with the appropriate plugins, Microsoft Office or OpenOffice.org. Moreover, there are numerous export filters to copy selected entries to websites, other formats, or tools to allow rapid sharing of findings and easy import into survey writing software.

Citation Network Statistics & Visualization

Once analysts have reviewed the data using standard reference management techniques, they can view visualizations of the citation network of the papers in the SocialAction network analysis tool (Perer & Shneiderman, 2006) (Fig. 2 (5–6)). Using these visualizations of the citation network we can easily find unexpected trends, clusters, gaps and outliers. Additionally, visualizations can immediately identify invalid data that is easily missed in

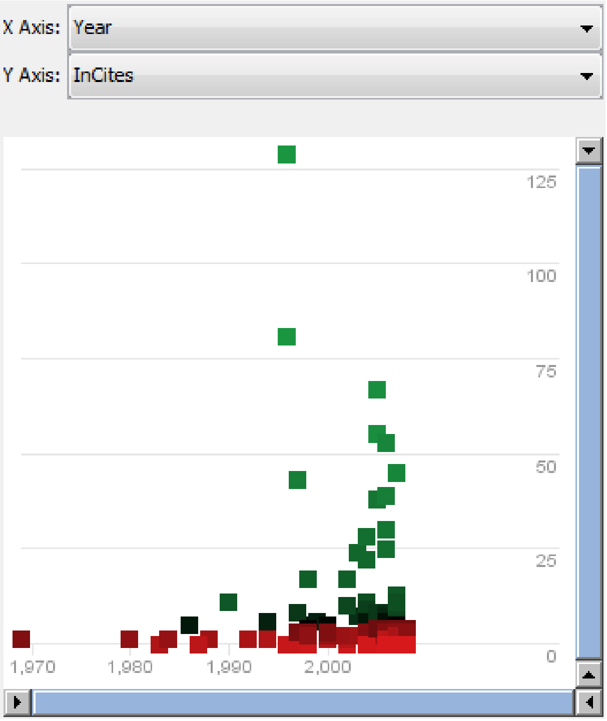


Figure 3. : This scatterplot shows paper publication year on the horizontal axis and the number of incoming citations on the vertical axis. There is a large spike in well-cited papers around 2005, two highly-cited outliers from 1996, and one poorly cited outlier in 1969.

tabular views.

The left view (Fig. 2 (5)) shows a ranking of papers by dynamically computed network statistics such as their in-degree, which is the number of citations to that paper within this dataset. Additional statistics include betweenness centrality, clustering coefficient, hubs or authorities, and any numeric attributes of the papers like year or externally computed measures. This ranked list can be filtered using the double-ended slider at its bottom, removing the top- or bottom-ranked papers in the list dynamically from the visualizations.

The papers in the collection can be viewed in standard charts like scatterplots (Fig. 3) to see trends and outliers, but visualizations of the network topology are more suited to finding research communities and tracking evolution over time. The node-link diagram of the network (Fig. 2 (6)) shows papers as rounded rectangle nodes, colored by their statistic rankings and connected by their citations using spline arrows. The nodes are arranged using a force-directed layout algorithm such that tightly connected nodes are placed in proximity to each other while loosely connected ones move to the extremes. As users filter or group nodes in the visible network the layout algorithm continues to run, updating the layout to reflect any changes. Nodes can also be colored by categorical attributes and users can compare nodes using scatterplots of their statistics (not shown). Edges can also be colored using statistical rankings, such as edge betweenness centrality or externally computed measures like citation sentiment analysis.

Papers can be grouped manually or using Newman's fast community-finding heuristic (Newman, 2004), which finds groups of papers that tend to cite each other more often than external papers. The found communities are shown using colored convex hulls surrounding the group, and the inter-group spring coefficients for the force-directed layout are reduced to separate them visually. Community-finding algorithms are most useful when exploring large datasets, though there are at least two meaningful communities shown in our examples, discussed below in Scenario: Dependency Parsing.

Citation Context

The node-link diagram shows users the number of citations to a paper and topology patterns, but it can also be useful to examine the context in which each of those citations were made in the citing paper. The context of citations often includes detailed and descriptive statements about the cited paper (Garfield, 1994) such as a summary, the paper's critical reception, and citations to follow-up papers (Giles, 1998; Bollacker, 1998).

From the full text of each paper ASE extracts the sentences containing the citations and their locations in the paper. Then, for any selected papers of interest, the context sentences of all citations to them are displayed in the citation context/in-cite text view (Fig. 2 (7)). If several papers are selected, all their context sentences are shown. Each sentence is a hyperlink that, when clicked, displays the full text of its source paper with the citation highlighed in the full text/out-cite text view (Fig. 2 (9)). Users can then see the broader context of the citation when the citing sentence alone is not sufficient.

Moreover, each citation in the full text is colored and hyperlinked to the target papers, allowing users to rapidly view the cited papers' metadata, full text, statistics, and network location while reading. The hyperlinks also provides immediate access to any cited follow-up papers. As we do not yet have character offsets within sentences for each citation, the entire sentence is hyperlinked to one of the citations. Any other citations found are hyperlinked to indices at the end of the sentence (e.g., the three additional citations represented in Fig. 2 (9) as D(#2 #3 #4)).

Multi-Document Summarization

Viewing the citation context for a paper or its abstract and keywords can give users an idea of its contribution to the field. However, highly-cited papers have too many citations to read through them all (see Fig. 2 (7) for an example). Furthermore, when looking at multiple papers selected manually or through the community-finding algorithm it can be difficult to understand the group's key focuses and various contributions.

To aid users in these tasks we provide automatically generated multi-document summaries for any selected set of papers, shown in Fig. 2 (8). Summaries of the full text of papers can be useful, but citation contexts and abstracts are richer in survey-worthy information. Mohammad (2009) shows that summaries based on citation contexts contain crucial survey-worthy information that is not available or hard to extract from abstracts and the full texts of papers. Likewise, they demonstrate that abstract summaries contain information not present in citation contexts and full texts.

Among the four summarization techniques compared by Mohammad (2009), the best at capturing the contributions of papers was Multi-Document Trimmer (MDT) (Zajic, Dorr, Schwartz, Monz, & Lin, 2005; Zajic, Dorr, Lin, & Schwartz, 2007), originally designed to summarize news articles. MDT is an extension of the original Trimmer which summarized single news articles (Dorr, Zajic, & Schwartz, 2003; Zajic, Dorr, & Schwartz, 2004). ASE uses MDT to provide summaries of citation context, but because citation sentences are not connected to each other and have metadata inline, we made some modifications to better handle this data. The only change to MDT used for the summaries in our examples was to remove inline metadata before processing, but we are currently exploring more fundamental refinements. For these examples we will focus on citation context summaries instead of using abstracts or full text, though ASE is modular in design and supports showing multiple summaries simultaneously. We will also show only multi-document summaries, but single-document citation context summaries using Trimmer or Cluster-Lexrank (Qazvinian & Radev, 2008) for highly-cited papers could be easily added.

Linking the Views

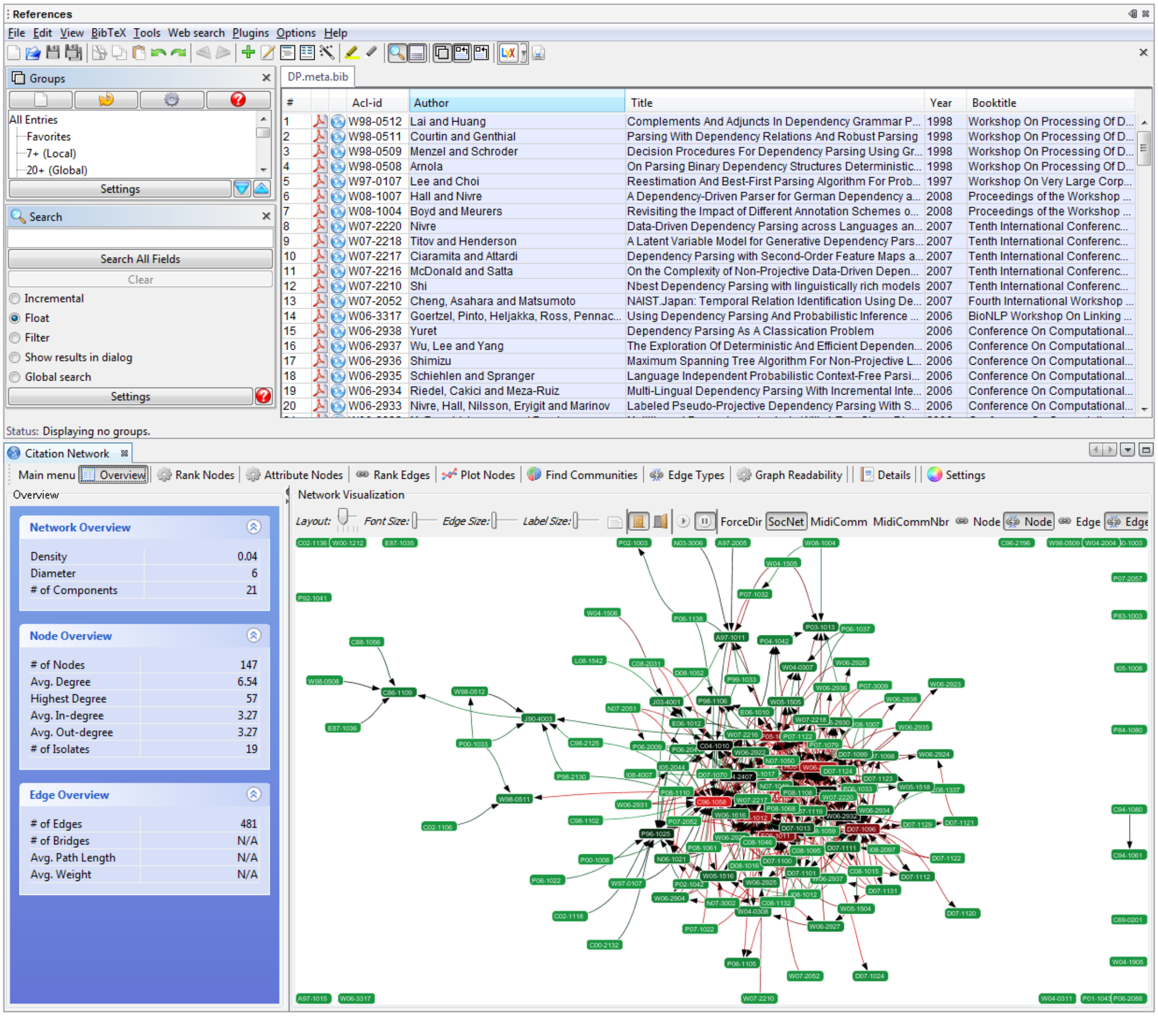
Each window presents a distinct view of the underlying scientific literature, each with its own advantages and disadvantages. While seeing paper metadata and opening the full text is easiest from the reference management view, determining the relationships between them is best done with the network visualization. Each of the data views becomes more powerful when they are tightly coupled together, such that interactions in one are visually reflected in the others. This technique is called *multiple coordinated views* (Shneiderman, 2009; North & Shneiderman, 1997).

Each of the views in ASE are linked to all the other windows. When users select papers in the reference manager the selection is also highlighted in the citation network visualization and the statistics ranked list. Likewise, the detail views show the papers' abstracts, reviews, reference forms, citation context, and generated summaries. Selecting nodes in the network visualization or any other view performs similarly, highlighting the nodes in all other views and showing their details.

The only exception to this linking is the full text view, which has two planned use cases. Once users bring up the full text view, they may wish to click on the hyperlinked citations within the text as they read. Clicking a citation selects the cited node in each of the other views, but to prevent the user from losing their place does not update the full text view.

The other use for the full text view is to only update when users select a citation in the citation context view to see the surrounding context in the citing paper's full text. We display which mode is currently being used by updating the border color of the view to show how it is currently interacting with others. Green, as seen in Fig. 2 (9), indicates that the full text view is showing the citation context for a selected citation. Blue, on the other hand, means that a citation within the full text has been selected, highlighting the cited paper in each other view.

In some situations screen space may be limited or users may wish to focus on a subset of the views. ASE provides a docking window manager interface that allows users to hide individual windows, resize or rearrange them, or even drag them to separate monitors. Revealing additional views to users as they gain experience can help reduce the initial learning curve faced when confronted with the entire ASE interface.



○ Incremental
● Float
○ Filter
○ Show results in dialog
○ Global search

#	Acl-id	Author	Title	Year	Booktitle
1	W98-0512	Lai and Huang	Complements And Adjuncts In Dependency Grammar P...	1998	Workshop On Processing Of D...
2	W98-0511	Courtin and Genthial	Parsing With Dependency Relations And Robust Parsing	1998	Workshop On Processing Of D...
3	W98-0509	Menzel and Schroder	Decision Procedures For Dependency Parsing Using Gr...	1998	Workshop On Processing Of D...
4	W98-0508	Arnola	On Parsing Binary Dependency Structures Deterministic...	1998	Workshop On Processing Of D...
5	W97-0107	Lee and Choi	Reestimation And Best-First Parsing Algorithm For Prob...	1997	Workshop On Very Large Corp...
6	W08-1007	Hall and Nivre	A Dependency-Driven Parser for German Dependency a...	2008	Proceedings of the Workshop ...
7	W08-1004	Boyd and Meurers	Revisiting the Impact of Different Annotation Schemes o...	2008	Proceedings of the Workshop ...
8	W07-2220	Nivre	Data-Driven Dependency Parsing across Languages an...	2007	Tenth International Conferenc...
9	W07-2218	Titov and Henderson	A Latent Variable Model for Generative Dependency Pars...	2007	Tenth International Conferenc...
10	W07-2217	Ciaramita and Attardi	Dependency Parsing with Second-Order Feature Maps a...	2007	Tenth International Conferenc...
11	W07-2216	McDonald and Satta	On the Complexity of Non-Projective Data-Driven Depen...	2007	Tenth International Conferenc...
12	W07-2210	Shi	Nbest Dependency Parsing with linguistically rich models	2007	Tenth International Conferenc...
13	W07-2052	Cheng, Asahara and Matsumoto	NAIST.Japan: Temporal Relation Identification Using De...	2007	Fourth International Workshop ...
14	W06-3317	Goertzel, Pinto, Heljakka, Ross, Pennac...	Using Dependency Parsing And Probabilistic Inference ...	2006	BioNLP Workshop On Linking ...
15	W06-2938	Yuret	Dependency Parsing As A Classication Problem	2006	Conference On Computational...
16	W06-2937	Wu, Lee and Yang	The Exploration Of Deterministic And Efficient Dependen...	2006	Conference On Computational...
17	W06-2936	Shimizu	Maximum Spanning Tree Algorithm For Non-Projective L...	2006	Conference On Computational...
18	W06-2935	Schiehlen and Spranger	Language Independent Probabilistic Context-Free Parsi...	2006	Conference On Computational...
19	W06-2934	Riedel, Cakici and Meza-Ruiz	Multi-Lingual Dependency Parsing With Incremental Inte...	2006	Conference On Computational...
20	W06-2933	Nivre, Hall, Nilsson, Eryigit and Marinov	Labeled Pseudo-Projective Dependency Parsing With S...	2006	Conference On Computational...

Status: Displaying no groups.

Network Overview	
Density	0.04
Diameter	6
# of Components	21

Node Overview	
# of Nodes	147
Avg. Degree	6.54
Highest Degree	57
Avg. In-degree	3.27
Avg. Out-degree	3.27
# of Isolates	19

Edge Overview	
# of Edges	481
# of Bridges	N/A
Avg. Path Length	N/A
Avg. Weight	N/A

Figure 4. : Starting interface with "Dependency Parsing" (DP) query loaded.

Limitations

The design of ASE also has many limitations that may undermine its advantages. The multiple windows require a large screen display to be useful and may increase perceptual and cognitive loads as users make selections which cause changes in multiple windows. Also, the integration of existing components means that there are different interfaces, especially in color highlighting, tool bars, and layouts. Moreover, the rich set of data extracted for each article means that preparing a collection is time consuming, thereby limiting our flexibility in conducting evaluations.

Scenario: Dependency Parsing

Imagine Karl, a student new to the field of Dependency Parsing (DP). DP is a small field of Computational Linguistics (CL) dedicated to analyzing sentences based on which of their components are dependent on each other. Karl first runs a search on the ACL

Anthology Network (AAN) for papers containing “Dependency Parsing”, which returns a subset of 147 papers and the citations between them. After loading the dataset ASE displays the initial windows shown in Fig. 4.

The top view of Fig. 4 shows Karl a reference management interface with a table of all the papers matching the search. In the bottom left he can see a statistical overview of the citation network, including the number of nodes and edges, average in- and out-degree of nodes, and the number of unconnected components. In the rest of the bottom half he can see the topology of the citation network in a node-link diagram, with individual papers colored by the number of citations they have received.

Identifying Key Papers & Authors

Karl is interested in identifying and reading the most influential papers in the field, so he clicks the “Rank Nodes” button to replace the overall statistics window with a list of papers ranked by their in-degree (Fig. 5). The in-degree of a paper is the number of citations it has received from other papers within this subset of the AAN. From here Karl selects all papers cited seven or more times (Fig. 6), and that subset is highlighted in the reference manager (top). He then drags these 14 papers to a group he created in the reference manager to keep track of those results (top left).

Karl quickly notices several things by scanning the table of these highly cited papers. First, all but four of the 14 are written by various combinations of the authors Nivre (6), Nilsson (6), and Hall (3) from Växjö University as well as McDonald (6) and Pereira (4) from University of Pennsylvania. Second, they are all written from 2004–2007, except for two written in 1996 separately by Eisner[2] (the most highly cited with 43) and Collins[3] (14). A simple search by author reveals that both Collins and Eisner have additional papers in the dataset, but only in the late 2000s and with few citations.

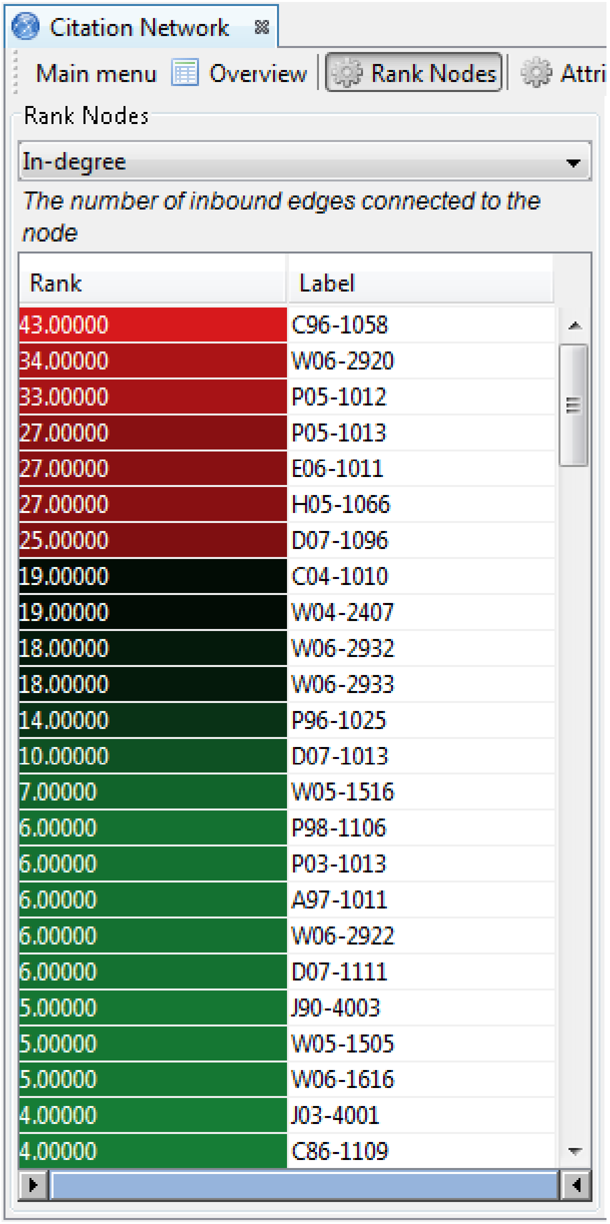
Karl thinks that he has seen the Collins paper cited before in another field of CL. To compare how many citations it has received among DP papers versus CL papers in general, he creates a scatterplot with those DP citations on the horizontal axis and CL citations on the vertical axis (Fig. 7). The selected Collins paper is shown with a white square, and it is immediately clear that it is highly cited in CL while less so in DP.

Karl then wants to see the citation network of only those highly cited papers, so he uses the double-ended slider at the bottom of the ranked list to filter out papers cited less than seven times. The filtered ranked list and citation network visualizations are shown in Fig. 8, and Karl can zoom into it or lay out only the filtered nodes to better see their citation patterns.

Now that Karl has stored a list of interesting papers he starts analyzing them in depth. For each one he selects, the citation context view displays the incoming citations for the paper. After selecting the key Eisner paper and scanning the incoming citations he finds one of particular interest to him: “Eisner (1996) introduced a data-driven dependency

[2] Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: an exploration. In *International conference on computational linguistics*. Retrieved from `http://clair.si.umich.edu/clair/anthology/query.cgi?type=Paper&id=C96-1058`.

[3] Collins, M. J. (1996). A new statistical parser based on bigram lexical dependencies. *Annual meeting of the association for computational linguistics*. Retrieved from `http://clair.si.umich.edu/clair/anthology/query.cgi?type=Paper&id=P96-1025`.



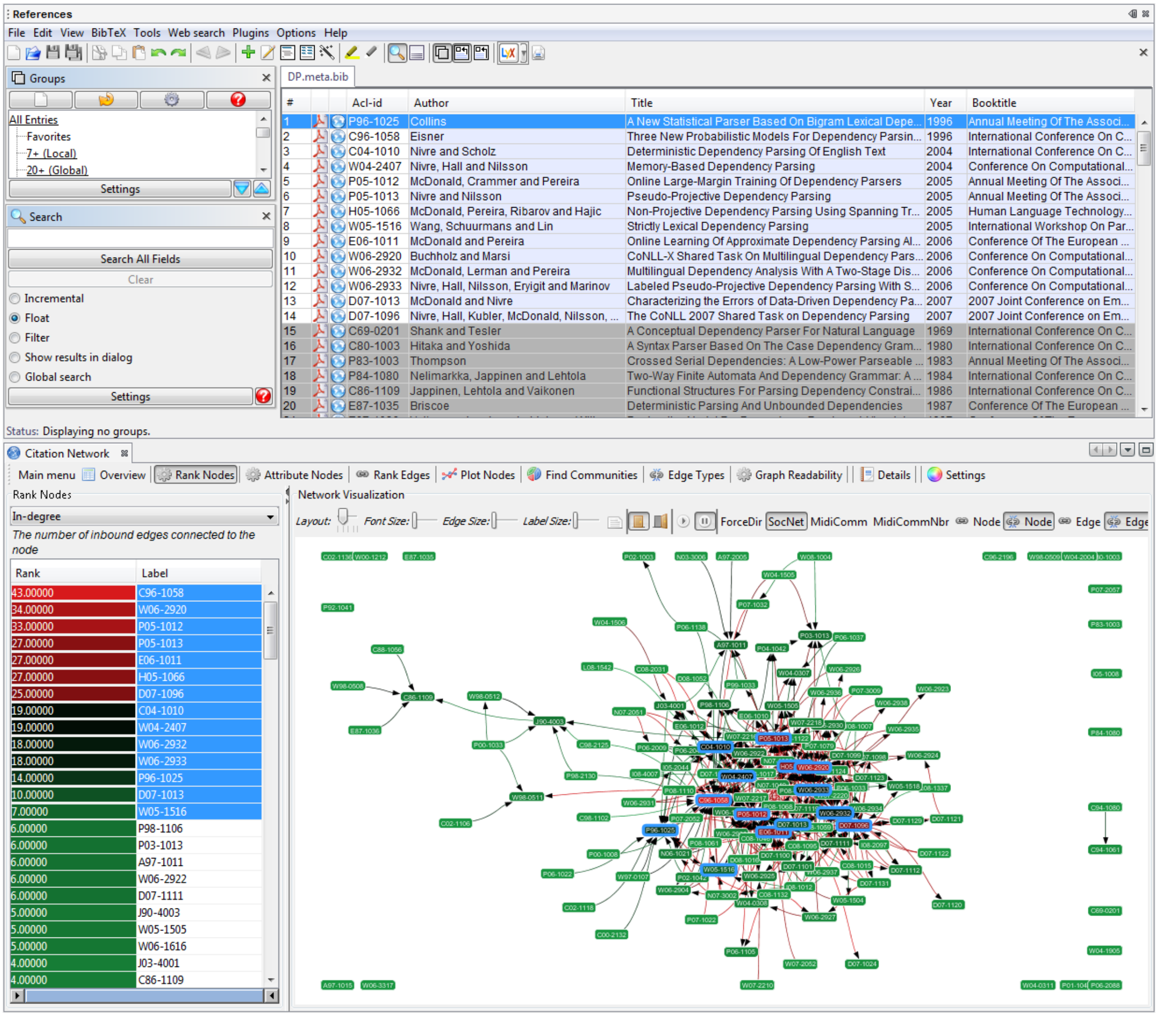
Rank Nodes

In-degree

The number of inbound edges connected to the node

Rank	Label
43.00000	C96-1058
34.00000	W06-2920
33.00000	P05-1012
27.00000	P05-1013
27.00000	E06-1011
27.00000	H05-1066
25.00000	D07-1096
19.00000	C04-1010
19.00000	W04-2407
18.00000	W06-2932
18.00000	W06-2933
14.00000	P96-1025
10.00000	D07-1013
7.00000	W05-1516
6.00000	P98-1106
6.00000	P03-1013
6.00000	A97-1011
6.00000	W06-2922
6.00000	D07-1111
5.00000	J90-4003
5.00000	W05-1505
5.00000	W06-1616
4.00000	J03-4001
4.00000	C86-1109

Figure 5. : Ranked list of DP papers by their in-degree (the citation count within this subset).



#	Acl-id	Author	Title	Year	Booktitle
1	P96-1025	Collins	A New Statistical Parser Based On Bigram Lexical Depe...	1996	Annual Meeting Of The Associ...
2	C96-1058	Eisner	Three New Probabilistic Models For Dependency Parsin...	1996	International Conference On C...
3	C04-1010	Nivre and Scholz	Deterministic Dependency Parsing Of English Text	2004	International Conference On C...
4	W04-2407	Nivre, Hall and Nilsson	Memory-Based Dependency Parsing	2004	Conference On Computational...
5	P05-1012	McDonald, Crammer and Pereira	Online Large-Margin Training Of Dependency Parsers	2005	Annual Meeting Of The Associ...
6	P05-1013	Nivre and Nilsson	Pseudo-Projective Dependency Parsing	2005	Annual Meeting Of The Associ...
7	H05-1066	McDonald, Pereira, Ribarov and Hajic	Non-Projective Dependency Parsing Using Spanning Tr...	2005	Human Language Technology...
8	W05-1516	Wang, Schuurmans and Lin	Strictly Lexical Dependency Parsing	2005	International Workshop On Par...
9	E06-1011	McDonald and Pereira	Online Learning Of Approximate Dependency Parsing Al...	2006	Conference Of The European ...
10	W06-2920	Buchholz and Marsi	CoNLL-X Shared Task On Multilingual Dependency Pars...	2006	Conference On Computational...
11	W06-2932	McDonald, Lerman and Pereira	Multilingual Dependency Analysis With A Two-Stage Dis...	2006	Conference On Computational...
12	W06-2933	Nivre, Hall, Nilsson, Eryigit and Marinov	Labeled Pseudo-Projective Dependency Parsing With S...	2006	Conference On Computational...
13	D07-1013	McDonald and Nivre	Characterizing the Errors of Data-Driven Dependency Pa...	2007	2007 Joint Conference on Em...
14	D07-1096	Nivre, Hall, Kubler, McDonald, Nilsson, ...	The CoNLL 2007 Shared Task on Dependency Parsing	2007	2007 Joint Conference on Em...
15	C69-0201	Shank and Tesler	A Conceptual Dependency Parser For Natural Language	1969	International Conference On C...
16	C80-1003	Hitaka and Yoshida	A Syntax Parser Based On The Case Dependency Gram...	1980	International Conference On C...
17	P83-1003	Thompson	Crossed Serial Dependencies: A Low-Power Parseable ...	1983	Annual Meeting Of The Associ...
18	P84-1080	Nelimarkka, Jappinen and Lehtola	Two-Way Finite Automata And Dependency Grammar: A ...	1984	International Conference On C...
19	C86-1109	Jappinen, Lehtola and Vaikonen	Functional Structures For Parsing Dependency Constrai...	1986	International Conference On C...
20	E87-1035	Briscoe	Deterministic Parsing And Unbounded Dependencies	1987	Conference Of The European ...

Status: Displaying no groups.

Rank Nodes

In-degree

The number of inbound edges connected to the node

Rank	Label
43.00000	C96-1058
34.00000	W06-2920
33.00000	P05-1012
27.00000	P05-1013
27.00000	E06-1011
27.00000	H05-1066
25.00000	D07-1096
19.00000	C04-1010
19.00000	W04-2407
18.00000	W06-2932
18.00000	W06-2933
14.00000	P96-1025
10.00000	D07-1013
7.00000	W05-1516
6.00000	P98-1106
6.00000	P03-1013
6.00000	A97-1011
6.00000	W06-2922
6.00000	D07-1111
5.00000	J90-4003
5.00000	W05-1505
5.00000	W06-1616
4.00000	J03-4001
4.00000	C86-1109

Figure 6. : DP papers with seven or more citations are highlighted in the ranked list (bottom-left), reference manager (top), and node-link diagram (bottom-right).

parser and compared several probability models on (English) Penn Treebank data" (Fig. 9, bottom). When he clicks on that citation, its surrounding context is displayed in the full text of the citing paper (top). From here he starts exploring the other hyperlink citations from that paper.

Finally, Karl can view the abstracts for each of those papers and open their full text in his PDF viewer to analyze them in depth. Throughout this process he takes notes in the review field of the reference manager to keep track of his insights.

Tracing the Topic Evolution

Now that Karl has an understanding of the key topics, he wants to trace the evolution of the topic over time. Similar to before, he ranks the papers by the year they were published and uses the double-ended slider to filter out all but the earliest year in the dataset. Then, by slowly dragging the right end of the slider he reveals the papers in the order they were

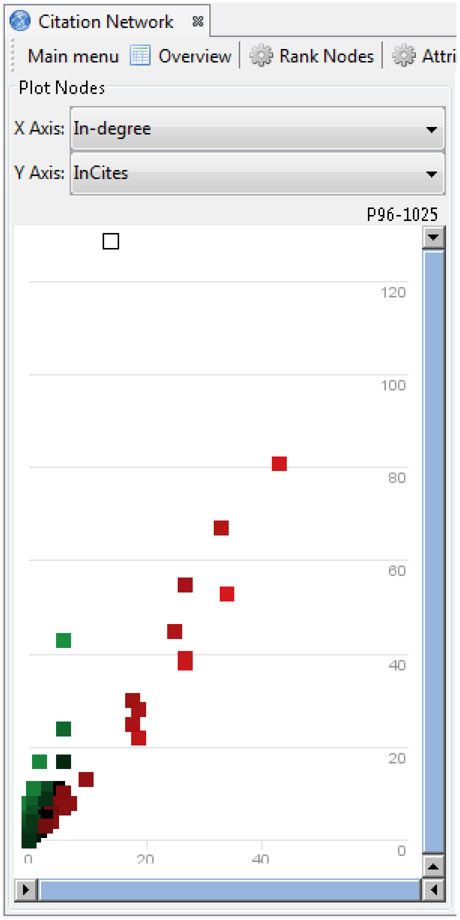


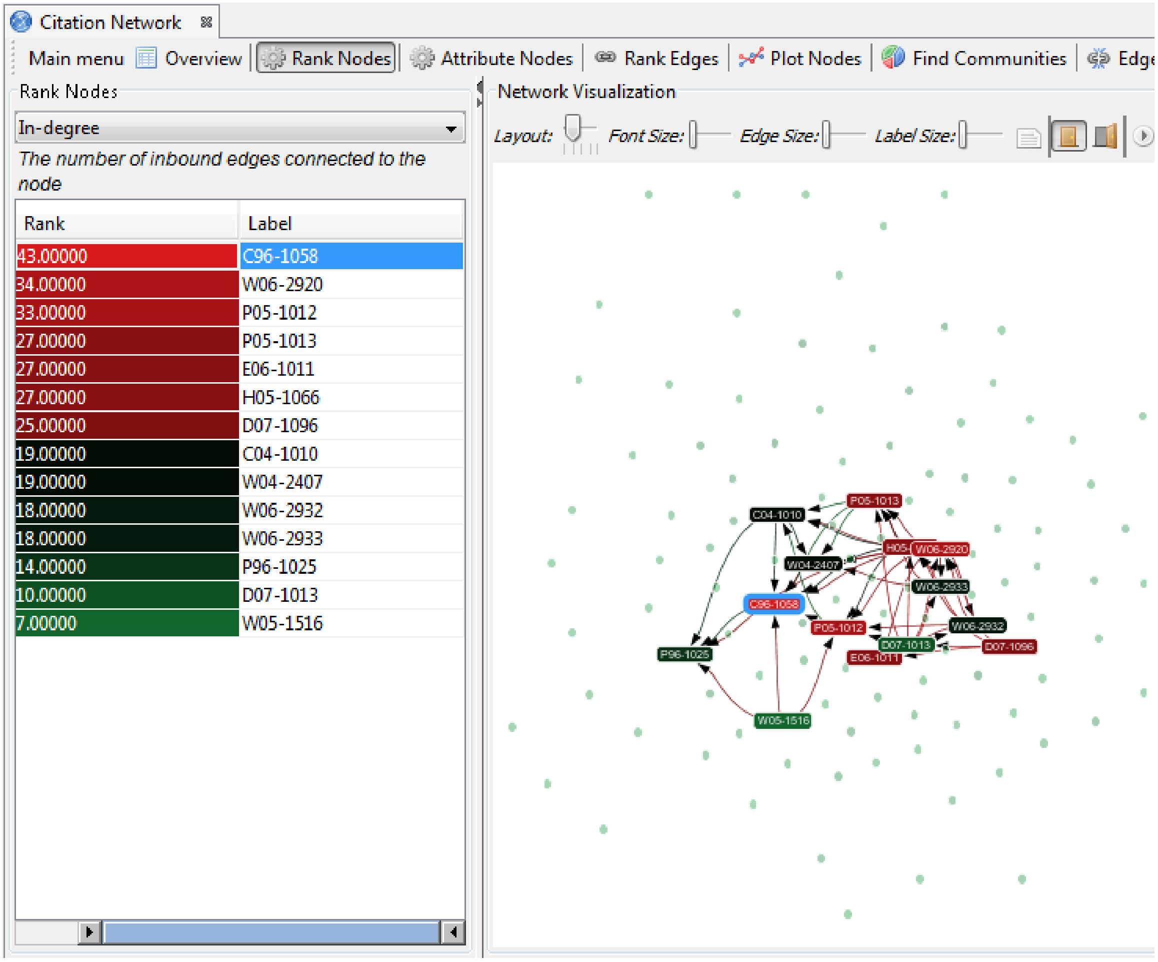
Figure 7. : This scatterplot has the number of citations within this subset on the horizontal axis and the number overall on the vertical axis. There is a general linear trend. The white box at the top shows the selected paper by Collins, which is highly cited in CL but less so in the DP subset.

published and the citations between them. He sees the first connected group of papers appearing from 1986–1998, seen in Fig. 10 (left). By CTRL-clicking on each paper, he displays them them in a table in the reference manager and discovers that they center around a research group from the SITRA Foundation in Helsinki, Finland[4].

However, after dragging the slider further Karl sees few papers connected to them in the following years. Starting in 1996, a disconnected group appears beginning with the highly cited Eisner and Collins papers he found in the previous section, which can be seen in the right side of Fig. 10. After filtering up to 1998, two papers (duplicates) by Lombardo and Lesmo[5] appear and cite both the SITRA and Eisner/Collins research communities.

[4] Jappinen, H., Lehtola, A., and Valkonen, K. (1986). Functional structures for parsing dependency constraints. *International conference on computational linguistics*. Retrieved from http://clair.si.umich.edu/clair/anthology/query.cgi?type=Paper&id=C86-1109.

[5] Lombardo, V. and Lesmo, L. (1998). Formal aspects and parsing issues of dependency theory. *Annual meeting of the association for computational linguistics and international conference on computational lin-*



In-degree

The number of inbound edges connected to the node

Rank	Label
43.00000	C96-1058
34.00000	W06-2920
33.00000	P05-1012
27.00000	P05-1013
27.00000	E06-1011
27.00000	H05-1066
25.00000	D07-1096
19.00000	C04-1010
19.00000	W04-2407
18.00000	W06-2932
18.00000	W06-2933
14.00000	P96-1025
10.00000	D07-1013
7.00000	W05-1516

Figure 8. : The ranked list and node-link diagram show only the papers cited more than seven times, filtered using the double-ended slider at the bottom-left.

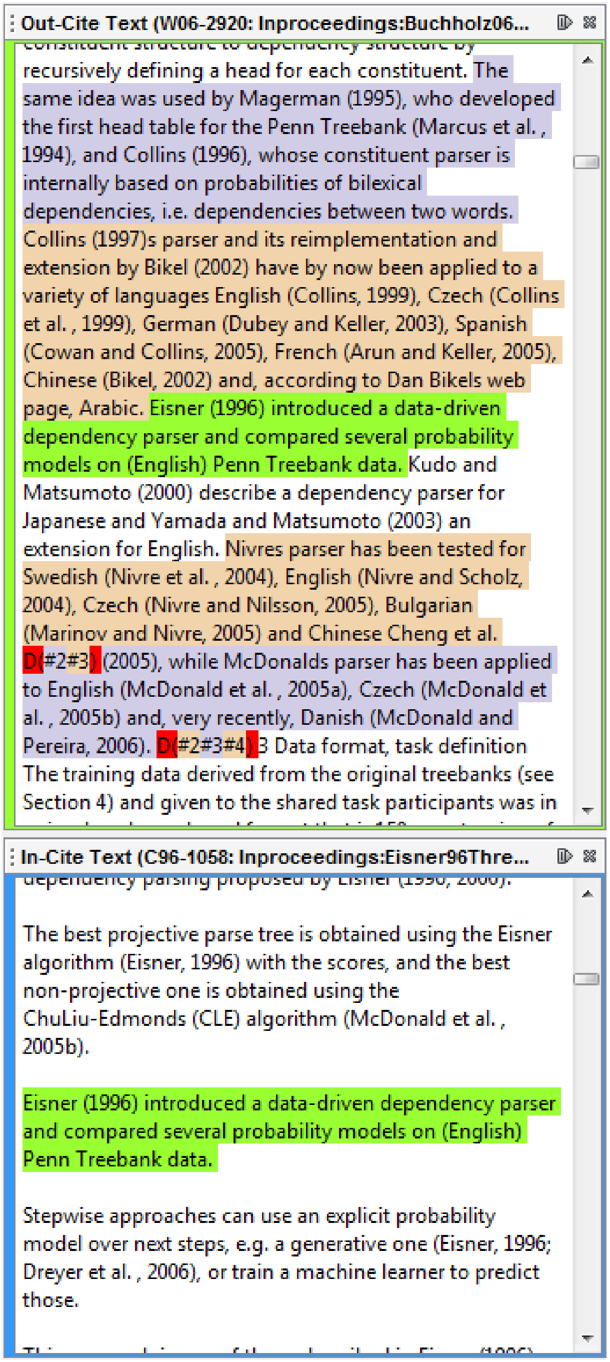
Continuing on, Karl finds that the vast majority of later work in DP is built around the later Eisner/Collins community with few citations to the SITRA group.

During 2006–2008 Karl sees an explosion in research on DP, with approximately 30 papers each year. Sorting the papers in the reference manager by year and scanning their venue, he finds that the bulk of the papers come from the 2006 and 2007 Conference on Computational Natural Language Learning (CoNLL) which both addressed DP.

Exploring Research Communities

As part of the topic evolution analysis Karl found two separate research communities using the force-directed layout and filtering. To more effectively find other communities of interest he decides to use the community-finding algorithm built into ASE. The groups of related papers are surrounded by convex colored hulls (Fig. 11), and he quickly spots the

guistics. Retrieved from http://clair.si.umich.edu/clair/anthology/query.cgi?type=Paper&id=P98-2130.



constituent structure to dependency structure by recursively defining a head for each constituent. The same idea was used by Magerman (1995), who developed the first head table for the Penn Treebank (Marcus et al. , 1994), and Collins (1996), whose constituent parser is internally based on probabilities of bilexical dependencies, i.e. dependencies between two words. Collins (1997)s parser and its reimplementation and extension by Bikel (2002) have by now been applied to a variety of languages English (Collins, 1999), Czech (Collins et al. , 1999), German (Dubey and Keller, 2003), Spanish (Cowan and Collins, 2005), French (Arun and Keller, 2005), Chinese (Bikel, 2002) and, according to Dan Bikels web page, Arabic. Eisner (1996) introduced a data-driven dependency parser and compared several probability models on (English) Penn Treebank data. Kudo and Matsumoto (2000) describe a dependency parser for Japanese and Yamada and Matsumoto (2003) an extension for English. Nivres parser has been tested for Swedish (Nivre et al. , 2004), English (Nivre and Scholz, 2004), Czech (Nivre and Nilsson, 2005), Bulgarian (Marinov and Nivre, 2005) and Chinese Cheng et al. D(#2#3) (2005), while McDonalds parser has been applied to English (McDonald et al. , 2005a), Czech (McDonald et al. , 2005b) and, very recently, Danish (McDonald and Pereira, 2006). D(#2#3#4) 3 Data format, task definition The training data derived from the original treebanks (see Section 4) and given to the shared task participants was in

dependency parsing proposed by Eisner (1996, 2000).

The best projective parse tree is obtained using the Eisner algorithm (Eisner, 1996) with the scores, and the best non-projective one is obtained using the ChuLiu-Edmonds (CLE) algorithm (McDonald et al. , 2005b).

Eisner (1996) introduced a data-driven dependency parser and compared several probability models on (English) Penn Treebank data.

Stepwise approaches can use an explicit probability model over next steps, e.g. a generative one (Eisner, 1996; Dreyer et al. , 2006), or train a machine learner to predict those.

Figure 9. : The citation context for the Eisner paper is shown in the bottom, and the context for the green selected citation is shown above in the full text view.

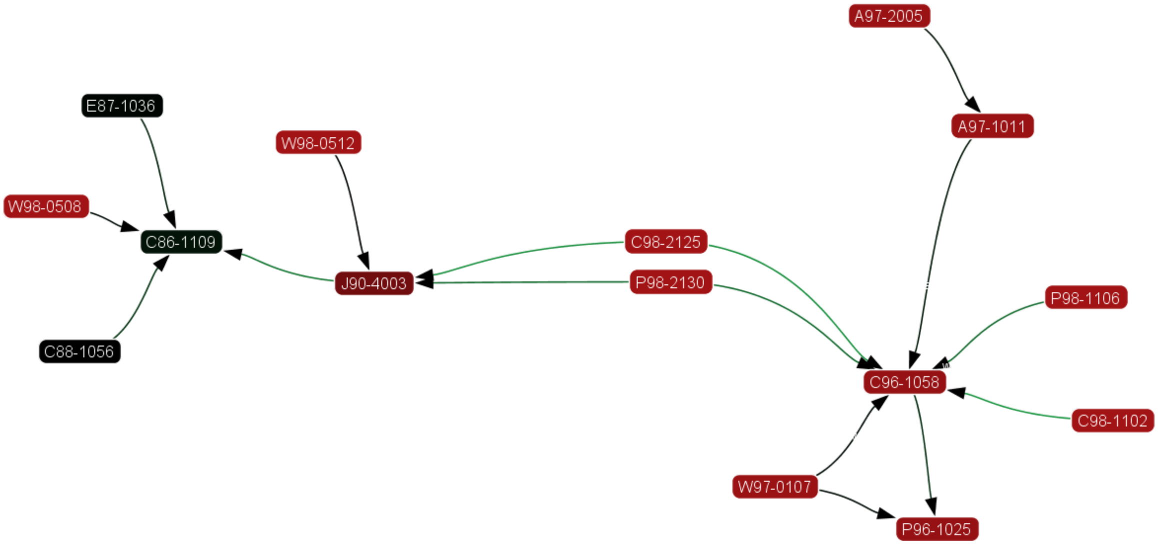
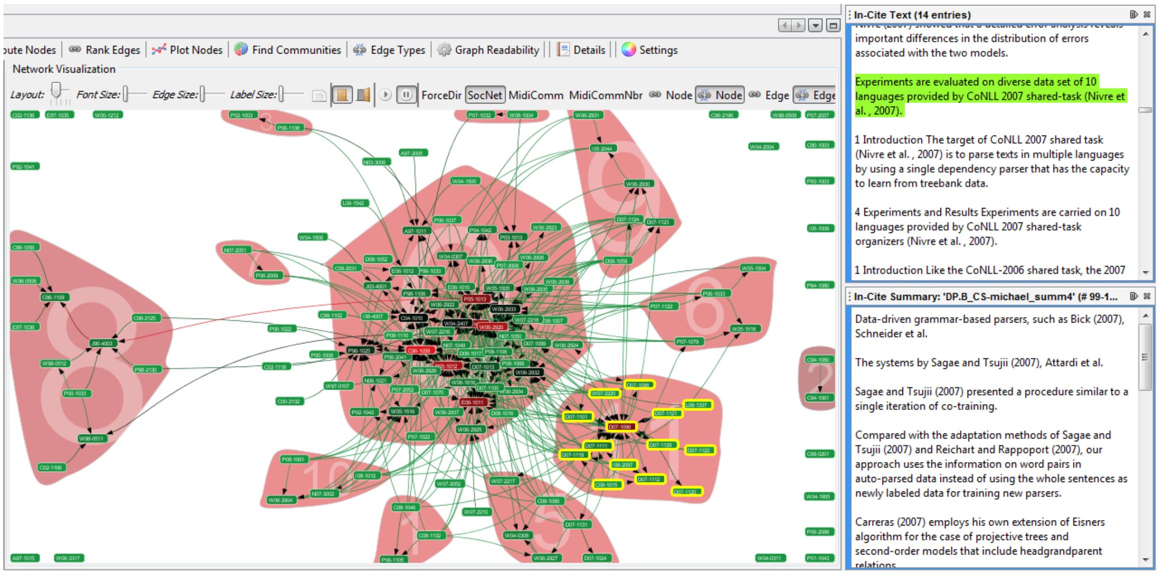


Figure 10. : The connected papers up through 1998 show the original 1986 community on the left and the new 1996 community growing on the right. They are bridged by two duplicate papers in 1998. By filtering to include subsequent years there is an explosion of research focused around the second, right community.



In-Cite Text (14 entries)

important differences in the distribution of errors associated with the two models.

Experiments are evaluated on diverse data set of 10 languages provided by CoNLL 2007 shared-task (Nivre et al. , 2007).

1 Introduction The target of CoNLL 2007 shared task (Nivre et al. , 2007) is to parse texts in multiple languages by using a single dependency parser that has the capacity to learn from treebank data.

4 Experiments and Results Experiments are carried on 10 languages provided by CoNLL 2007 shared-task organizers (Nivre et al. , 2007).

1 Introduction Like the CoNLL-2006 shared task, the 2007

In-Cite Summary: 'DP.B_CS-michael_summ4' (# 99-1...

Data-driven grammar-based parsers, such as Bick (2007), Schneider et al.

The systems by Sagae and Tsujii (2007), Attardi et al.

Sagae and Tsujii (2007) presented a procedure similar to a single iteration of co-training.

Compared with the adaptation methods of Sagae and Tsujii (2007) and Reichart and Rappoport (2007), our approach uses the information on word pairs in auto-parsed data instead of using the whole sentences as newly labeled data for training new parsers.

Carreras (2007) employs his own extension of Eisners algorithm for the case of projective trees and second-order models that include headgrandparent relations

Figure 11. : Algorithmically found communities are shown using convex hulls in the node-link diagram. When selected, all the citation context is shown in the top-right, along with an automatically generated summary of the overall context (bottom-right).

two groups he identified at the left and center.

However, the center core group was split by the community-finding algorithm into several smaller groups that were not obvious before. By clicking on the largest of these (bottom-right & highlighted in yellow), Karl sees the table of papers in it in the reference manager, all the citation context for the cluster (right), and an automatically generated summary of the citation context (bottom-right). He then scans the citations to these papers and sees frequent references to the CoNLL 2007 shared task that he saw before.

Zooming in on the community in the citation network to examine the citation edges, Karl notices that there are many unusual bi-directional citations between a central paper (Nivre[6]) and other papers in the cluster. By viewing the abstract of the Nivre paper Karl finds the reason for the bi-directional citations: these papers were written collaboratively. The Nivre paper provides an overview of the shared task for the year, the datasets used, and analyzes the differing approaches and results of the submitted systems. Karl reads through the citation context summary for a quick overview of the approaches of these papers. Later, he can dig deeper by reading the entire citation context or by viewing the full text of the Nivre paper.

Implementation Details

ASE is built using Java and the NetBeans Platform (Oracle, 2011) for window and settings management. The reference management view uses a version of the JabRef reference manager (JabRef Development Team, 2011) that was modified to interface with our brushing and linking framework. The citation network visualization and analysis components come from the SocialAction network analysis tool (Perer & Shneiderman, 2006), which was similarly altered to integrate our framework and automated loading of datasets. The remaining views in the interface for the citation context, automatically generated summaries, and full text are built using standard Java Swing widgets.

Data Import

Each of the search results from the ACL Anthology Network (AAN) (Radev, 2009; Radev, 2009) includes the paper metadata, abstract, citation context, and plain text. Initial loading of search results from the AAN is done by processing the results to create the standard data files used by JabRef and SocialAction, BibTeX and the HCIL Network Visualization Input Data Format[7], respectively. Each of the paper entries was modified to include unigram and bigram keywords generated from the plain text, a link to the AAN website for that paper, and an automatically downloaded full text PDF.

Multi-Document Summarization

For multi-document summarization we use a modified version Multi-Document Trimmer (MDT) (see the Design section on *Multi-Document Summarization*). Our current implementation of MDT processes the full text of each document in a selected group and

[6] Nivre, J., Hall, J., Kubler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. *2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. Retrieved from http://clair.si.umich.edu/clair/anthology/query.cgi?type=Paper&id=D07-1096.

[7] http://www.cs.umd.edu/hcil/nvss/netFormat.shtml

requires substantial computing time to build some of the summaries. To provide interactive response times for users, we pre-computed summaries for each of the communities output by Newman's fast community-finding heuristic (Newman, 2004) at each of its cutoff thresholds. These summaries are saved with the metadata and accessed when users select individual communities.

The summarization process could be sped up and generalized to arbitrary group selections by pre-computing sentence candidates for the selected texts of each paper (citation context, abstract, or full text). The pre-computation would use the syntactic trimming and shortening initial steps of MDT. Then, when users select communities or other groups of interest, the candidate sentences would be scored for relevance to the selected set and chosen based on their features using the remaining steps of MDT. With this optimization, the summarization time for each cluster could be reduced by a few minutes. Whether this optimized approach would be suitable for real-time summarization is an interesting next step.

Brushing, Linking, & Context

ASE provides a modular brushing and linking framework that allows any view to register itself to send updates to other views, as well as receive updates sent by others. A view does this by implementing the standard Java `PropertyChangeListener` (`PCL`) interface so it can receive updates, then registering itself with a central class for the kinds of updates it wants to send and receive (e.g., `BRUSHING`, `LINKING`, and `CONTEXT` selection). The central class coordinates these registrations, adding the view as a listener of the central class for the event types it can receive. Likewise, the central class implements `PCL` and adds itself as a listener for the view for each event type that view can send. When the central class receives a `PropertyChangeEvent` (`PCE`) from any of the views, it fires a property change itself to forward the event to any relevant views.

For example, when a selection is made in the reference manager view it fires a `LINKING` property change that includes all the selected BibTeX entries. As the central class is a listener for `LINKING` property changes, it receives the `PCE` and fires it off to the views registered to receive them. The network analysis view receives the `LINKING PCE`, extracts the unique ID from the sent BibTeX entries, and selects the associated nodes in each of its visualizations. The citation context view takes each of the sent BibTeX entries, loads the associated citation context data, and displays the concatenated result. The reference manager receives and ignores the `LINKING PCE` because it was the sender.

Similar `LINKING` processes occur when selections are made in any of the other views. Likewise, when items are moused over in a view, a `BRUSHING` event for those BibTeX entries is sent out so other views highlight the same items. The `CONTEXT` events are used to coordinate citation sentence selection in the citation context and full text views. Instead of plain BibTeX entries, these events contain citation line wrappers for the BibTeX entries and the citation metadata (e.g., source, target, text, and offsets in the source document). This discussion has omitted many details of the data structure sent with each PCE.

Evaluation

We conducted a planned, iterative user study procedure with refinements along the way to both ASE and our testing methods. The evaluation consisted mainly of three qualitative usability studies over 17 months to evaluate the effectiveness of ASE for exploring collections of papers. An early formative study with five participants helped identifying usability issues, guided the development of ASE, and determinted the tasks users were interested in performing with the tool. This helped us plan two subsequent and more structured usability studies.

For all three evaluations we used the same Dependency Parsing dataset described in the Design section *Search & Data Import* and Scenario: Dependency Parsing. It is important to have user study participants analyze data of interest to them, and preferably their own data, to keep them motivated and to give the tool significance (Plaisant, 2004). Thus, we recruited researchers interested in and knowledgeable about Computational Linguistics as our participants for each study.

Here we will focus on a high-level overview of the studies and their results without delving into their details. Highly detailed descriptions of the studies and the results of each participant are described for the second study in Gove, Dunne, Shneiderman, Klavans, and Dorr (2011) and for both the second and third studies in Gove (2011).

Second Study

Our second study was designed to evaluate the usability and effectiveness of ASE after refining both the tool and testing methods during the formative evaluation.

Participants. There were four participants in the second study: two current Computer Science PhD students and two recent graduates. Of these, two had prior experience with Dependency Parsing.

Procedure. The ASE evaluations were conducted using a 30-inch LCD monitor with a resolution of 1920x1080, running off an Intel Core i3 2.26 Ghz laptop with 4 GB of RAM. Sessions were limited to 120 minutes, starting with a 30 minute training session. For the training phase we showed the participants video clips demonstrating each of the features of ASE. Between videos, we asked them to practice the tasks shown and ask questions if they did not understand the tool or its features.

We provided participants with two predefined tasks determined via our formative studies, taking around 60 minutes to complete. We asked participants to: (1) identify and make note of important authors and papers, and (2) find an important paper and collect evidence to determine why it is important. These open-ended tasks allowed participants to use whatever features of the tool they thought would be useful, while providing a basic benchmark for their performance.

For the remaining 30 minutes, we asked them to identify additional tasks of interest to them using the dataset. From these we selected one or more as individual goals for the remainder of the session and asked the participant to try to perform them using ASE.

Throughout the study we asked participants to use a think-aloud approach, making note of their thoughts and actions. We made note of which capabilities demonstrated in the training videos were used by each participant, for both the predefined and individual tasks. At the conclusion of the session participants were asked to comment on their experiences using the system.

Results. The second study demonstrated that users were able to quickly grasp the basics of the reference manager and network visualization views, with some even using more advanced features of them. The overall view available in the node-link diagram was used frequently by participants to orient themselves, as well as to find interesting clusters, trends, and motifs in the topology. This illustrated the value of using multiple coordinated views to provide an overview of the dataset. Unfortunately, there seemed to be a steep learning curve with participants using the same set features at the beginning of the session as at the end.

By far the most used feature was ranking and filtering by paper metadata or computed network statistics. As the predefined tasks focused on finding "important" papers and authors, perhaps the participants found the provided rankings by quantitative measures to be easy jumping-off points. Similarly, filtering by a metric provides a quick drill-down to the "important" papers (according to that metric).

The participants showed great interest in the citation context view, scanning it for interesting papers, authors, and insights. However, they had problems analyzing more recent or other poorly-cited papers due to the little or no context available. Moreover, the interaction between citation context and the other views was challenging for one user who wanted to open the PDF of each citing paper without changing the visualization focus to them.

The interactions between the full text view and the other views were difficult for participants to understand, as each click on a citation changed the paper selected in all the other views but not in it. Perhaps a better indication of its relationships to the rest would be helpful, but this demonstrates once again that having systematic, homogeneous interactions across all views helps users understand the relationships.

While the participants were interested in exploring the multi-document summary feature, they were generally dissatisfied with the output quality of the summarization algorithm. MDT is designed to summarize news articles, and we found that citation sentences have several differences that need to be accounted for. For example, inline metadata and the disjoint nature of the sentences reduces the utility of MDT.

From the results of this study, we identified and implemented several improvements for the interactions between the views in ASE. Moreover, we adjusted the MDT summarization algorithm so as to better handle citation context instead of news articles.

Third Study

We ran our third user study approximately six months after the second one, both to test the usability problems found and fixed before and to evaluate the usage patterns of more experienced users.

Participants. The participants of the third study were four current Computer Science PhD students, two of which had participated in the second study as well. All four indicated some knowledge of the concept of Dependency Parsing, if not the associated literature.

Procedure. Our procedure for the third study was identical to the second, with the sole additions of screen and audio capture during the evaluation session for later analysis.

Results. The new participants confirmed our previous observations about the ease of use and value of the coordinated reference manager and network visualization views. Overall the participants used the same general approaches, including extensive use of the ranking and filtering features. However, the two repeat participants that were in the second study used more features their second time around and were able to find deeper insights in the dataset. This demonstrated the value of using extended duration evaluation techniques such as Multi-dimensional In-depth Long-term Case studies (MILCs) (Shneiderman & Plaisant, 2006), which focus on actual use of the system by domain experts solving their own problems. MILCS are well suited to evaluating creativity and exploration tools such as ASE that may be too complicated to understand in a single analysis session, though due to data import constraints we were unable to use a MILC study.

The improvements we applied to the MDT summarization algorithm and the interactions between views seemed to help users with their analyses. The new citation context summaries were used frequently during this study, and the participants were more satisfied with the linguistic structure of the summaries. They found that there were often coherent summaries of the themes in smaller communities, but were unable to find clear themes for larger ones. This is to be expected given the small size of this dataset and the large central community. Additionally, participants wanted more types of community summaries like topic modeling or using abstracts and full texts instead of citation context.

The automatic community finding algorithm was used by participants for several tasks, however it was limited by the small size of the dataset and by the types of communities it produced. Participants wanted additional clustering techniques for particular tasks and process models, and that were not limited to only clustering based on topology. Moreover, they wanted to select arbitrary sets of papers to summarize instead of being limited to the sets found by the clustering algorithm. This capability is limited by the speed the multi-document summarization algorithm. Unfortunately MDT is not fast enough for this currently, though the Implementation Details section on *Multi-Document Summarization* discusses one potential improvement.

Discussion & Future Work

From our three usability tests we found that users can quickly grasp the basics of the multiple coordinated views provided by ASE, though there is a steep learning curve for many of the features. However, our repeat participants demonstrated that with more sessions with the tool they can use more features and find deeper insights than they could initially. From the evaluations we discovered several usability issues with ASE, most of which we were able to correct and test again in the last user study. The improvements we made seemed to be effective, especially the coherence of the summaries generated by our modified version of MDT.

The user studies also helped us to identify several common questions users ask when exploring paper collections. Foremost they wanted to identify the foundations, breakthroughs, state-of-the-art, and evolution of a field. Next, they were looking to find collaborators and relationships between disparate communities. They were also searching for easily understandable overview papers like surveys to help guide their exploration.

We also developed a set of user requirements for exploring scientific literature networks to help guide the design process. First, users want control over the collection they are exploring. They want to choose a custom subset via a query and iteratively refine and drill down into it, putting them in control of the analysis. Next, users appreciate an overview of the subset either as a visualization or textual statistics. Overviews help users orient themselves in the subset and allow them to quickly browse via details-on-demand or other multiple coordinated view approaches. Our users made extensive use of the ranking and filtering features, demonstrating that easy to understand metrics for identifying interesting papers can provide a jump off point for more detailed analyses. Moreover, users should be able to create groups of papers and annotate them with their findings. Grouping and annotating helps users organize their discovery process, and lets them save their analyses and come back to them over a period of days or weeks.

Likewise, we identified several recommendations for future researchers conducting similar evaluations. Foremost, the use of extended user studies like MILCs (Shneiderman & Plaisant, 2006) is well suited for evaluating complex creativity and exploration tools like ASE. We were limited by our use of shorter evaluation sessions that only scratch the surface of available features, though our repeat participants were able to expand their repertoire and find deeper insights. One way to improve the tutorial retention is to follow the suggestions of Plaisant and Shneiderman (2005), which suggests having short clips about the functionality available throughout the sessions for participants to refresh their memory. Similarly, embedded training, animations, or slowly revealing features may help guide users in using the full capabilities of the system. Finally, the importance of motivating participants can not be stressed enough. Identify your target participants early and allow easy import from one or more general data sources of interest to them so they can analyze their own data.

These recommendations would have helped us, as our evaluation is limited by many of these issues. It is difficult to import new datasets into ASE due to the processing required. The collection we used contains only 147 papers, though in our evaluations participants were still able to find interesting insights. We had to select participants interested in the research area rather than letting them use their own datasets, which limited the pool of available researchers and their motivation. In the end, we only had six participants and were not able to recruit any for a longer MILC study. However, we still found many useful insights and usability fixes.

Conclusion

Understanding scientific domains and topics is a challenging task that is not well supported by current search systems. Fact, document, or exploratory search might require only minutes or hours to attain success, but understanding emerging research fields can take days or weeks. By integrating statistics, text analysis, and visualization we have some hope of providing users with the tools they need to generate readily-consumable surveys of scientific domains and topics. Our prototype implementation Action Science Explorer (ASE) combines reference management, statistics, citation context, automatic summarization, ranking & filtering, and network visualization in several coordinated views.

The three-phase usability study guided our revisions to ASE and led us to improve the testing methods. These evaluations demonstrated the utility of showing several coordinated views of a paper collection. Moreover, they identified several exploration tasks users are interested in and the benefit of specific functionality when performing them. The evaluations also found many limitations of ASE including the large screen space required and inconsistent user interfaces between views. The extensive data processing needed for the citation context views and summaries limits our data sources, though more and more academic databases are starting to integrating these features. Our multi-document summarization techniques produce useful summaries of citation context, though we are currently exploring refinements to better handle the disjoint nature of the citation text.

References

Aris, A., Shneiderman, B., Qazvinian, V., & Radev, D. (2009). Visual overviews for discovering key papers and influences across research fronts. *JASIST: Journal of the American Society for Information Science and Technology*, *60*(11), 2219–2228. doi:10.1002/asi.21160

Association for Computing Machinery. (2011). ACM Digital Library. Retrieved from http://portal.acm.org

Bates, M. J. (1990). Where should the person stop and the information search interface start? *Information Processing & Management*, *26*(5), 575–591. doi:10.1016/0306-4573(90)90103-9

Bollacker, K. D., Lawrence, S., & Giles, C. L. (1998). CiteSeer: an autonomous Web agent for automatic retrieval and identification of interesting publications. In *AGENTS '98: proc. second international conference on autonomous agents* (pp. 116–123). New York, NY, USA: ACM. doi:10.1145/280765.280786

Bradshaw, S. (2003). Reference directed indexing: redeeming relevance for subject search in citation indexes. In T. Koch & I. Slvberg (Eds.), *ECDL '03: proc. 7th european conference on research and advanced technology for digital libraries* (Vol. 2769, pp. 499–510). Lecture Notes in Computer Science. Springer Berlin / Heidelberg. doi:10.1007/978-3-540-45175-4_45

Center for History and New Media, G. (2011). Zotero [Software]. Retrieved from http://www.zotero.org

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *PNAS: Proc. National Academy of Sciences of the United States of America*, *101*(90001), 5303–5310. doi:10.1073/pnas.0307513100

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST: Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. doi:10.1002/asi.20317

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *JASIST: Journal of the American Society for Information Science and Technology*, *61*, 1386–1409. doi:10.1002/asi.v61:7

Cornell University Library. (2011). ArXiv. Retrieved from http://arxiv.org

Dorr, B., Zajic, D., & Schwartz, R. (2003). Hedge Trimmer: a parse-and-trim approach to headline generation. In *HLT/NAACL-DUC '03: proc. HLT/NAACL 2003 text summarization workshop and document understanding conference* (pp. 1–8). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1119467.1119468

Elsevier. (2011). SciVerse Scopus. Retrieved from http://scopus.com/

Garfield, E. (1994). The concept of citation indexing: a unique and innovative tool for navigating the research literature. *Current Contents*. Retrieved from http://thomsonreuters.com/products_services/science/free/essays/concept_of_citation_indexing/

Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). CiteSeer: an automatic citation indexing system. In *DL '98: proc. 3rd ACM conference on digital libraries* (pp. 89–98). New York, NY, USA: ACM. doi:10.1145/276675.276685

Google. (2011). Google Scholar. Retrieved from http://scholar.google.com

Gove, R. (2011). *Understanding scientific literature networks: case study evaluations of integrating vizualizations and statistics*. (Master's thesis, University of Maryland). Retrieved from http://www.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2011-11

Gove, R., Dunne, C., Shneiderman, B., Klavans, J., & Dorr, B. (2011). Evaluating visual and statistical exploration of scientific literature networks. In *VL/HCC '11: proc. 2011 IEEE symposium on visual languages and human-centric computing*. Retrieved from http://www.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2011-02

Hearst, M. A. (2009). *Search user interfaces* (1st). Cambridge University Press. Retrieved from http://searchuserinterfaces.com/book

Institute of Electrical and Electronics Engineers. (2011). IEEE Xplore. Retrieved from http://ieeexplore.ieee.org

JabRef Development Team. (2011). JabRef [Software]. Retrieved from http://jabref.sourceforge.net

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *JASIS: Journal of the American Society for Information Science*, *42*(5), 361–371. doi:10.1002/(SICI)1097-4571(199106)42:5%3C361::AID-ASI6%3E3.0.CO;2-%23

Marchionini, G. (1997). *Information seeking in electronic environments*. Cambridge Series on Human-Computer Interaction. Cambridge University Press. Retrieved from http://www.ils.unc.edu/~march/isee_book/web_page.html

Marchionini, G. (2006). Exploratory search: from finding to understanding. *CACM: Communications of the ACM*, *49*, 41–46. doi:10.1145/1121949.1121979

Mei, Q., & Zhai, C. (2008). Generating impact-based summaries for scientific literature. In *ACL/HLT '08: proc. 46th annual meeting of the association for computational linguistics: human language technologies* (pp. 816–824). Columbus, Ohio: Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology-new/P/P08/P08-1093

Mendeley Ltd. (2011). Mendeley [Software]. Retrieved from http://www.mendeley.com

Microsoft Research. (2011). Microsoft Academic Search. Retrieved from http://academic.research.microsoft.com

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., . . . Zajic, D. (2009). Using citations to generate surveys of scientific paradigms. In *HLT/NAACL '09: proc. human language technologies: the 2009 annual conference of the north american chapter of the association for computational linguistics* (pp. 584–592). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1620754.1620839

Nanba, H., Abekawa, T., Okumura, M., & Saito, S. (2004). Bilingual PRESRI: integration of multiple research paper databases. In C. Fluhr, G. Grefenstette & W. B. Croft (Eds.), *RIAO '04: proc. 7th international conference on computer-assisted information retrieval (recherche d'information assistee par ordinateur)* (pp. 195–211). Avignon, France: CID. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.102.9560

Nanba, H., & Okumura, M. (1999). Towards multi-paper summarization using reference information. In *IJCAI '99: proc. 16th international joint conference on artificial intelligence* (pp. 926–931). Retrieved from http://portal.acm.org/citation.cfm?id=1624312.1624351

National Center for Biotechnology Information. (2011). PubMed. Retrieved from http://ncbi.nlm.nih.gov/pubmed

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *PNAS: Proc. National Academy of Sciences of the United States of America*, *98*(2), 404–409. doi:10.1073/pnas.021544898

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, *69*(6), 066133. doi:10.1103/PhysRevE.69.066133

North, C., & Shneiderman, B. (1997). *A taxonomy of multiple window coordinations* (Human-Computer Interaction Lab Tech Report No. HCIL-97-18). Retrieved from http://www.cs.umd.edu/local-cgi-bin/hcil/rr.pl?number=97-18

NWB Team. (2006). Network Workbench [Software]. Retrieved from http://nwb.slis.indiana.edu

Oracle. (2011). NetBeans Platform [Software]. Retrieved from http://netbeans.org/features/platform

Perer, A., & Shneiderman, B. (2006). Balancing systematic and flexible exploration of social networks. *TVCG: IEEE Transactions on Visualization and Computer Graphics*, *12*(5), 693–700. doi:10.1109/TVCG.2006.122

Plaisant, C. (2004). The challenge of information visualization evaluation. In *AVI '04: proc. 2004 working conference on advanced visual interfaces* (pp. 109–116). New York, NY, USA: ACM. doi:10.1145/989863.989880

Plaisant, C., & Shneiderman, B. (2005). Show me! Guidelines for producing recorded demonstrations. *VLHCC '05: Proc. 2005 IEEE Symposium on Visual Languages and Human-Centric Computing*, *00*, 171–178. doi:10.1109/VLHCC.2005.57

Qazvinian, V., & Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *COLING '08: proc. 22nd international conference on computational linguistics* (pp. 689–696). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1599081.1599168

Radev, D., Otterbacher, J., Winkel, A., & Blair-Goldensohn, S. (2005). NewsInEssence: summarizing online news topics. *Communications of the ACM*, *48*, 95–98. doi:10.1145/1089107.1089111

Radev, D. R., Joseph, M. T., Gibson, B., & Muthukrishnan, P. (2009). A bibliometric and network analysis of the field of computational linguistics. *JASIST: Journal of the*

American Society for Information Science and Technology. To appear. Retrieved from http://clair.si.umich.edu/~radev/papers/biblio.pdf

Radev, D. R., Muthukrishnan, P., & Qazvinian, V. (2009). The ACL Anthology Network corpus. In *NLPIR4DL '09: proc. ACL-IJCNLP 2009 workshop on text and citation analysis for scholarly digital libraries* (pp. 54–61). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1699750.1699759

Sekine, S., & Nobata, C. (2003). A survey for multi-document summarization. In D. Radev & S. Teufel (Eds.), *HLT/NAACL-DUC '03: proc. HLT/NAACL 2003 text summarization workshop and document understanding conference* (pp. 65–72). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1119467.1119476

Shneiderman, B., & Plaisant, C. (2006). Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *BELIV '06: proc. 2006 avi workshop on BEyond time and errors: novel evaLuation methods for Information Visualization* (pp. 1–7). New York, NY, USA: ACM. doi:10.1145/1168149.1168158

Shneiderman, B., Plaisant, C., Cohen, M., & Jacobs, S. (2009). *Designing the user interface: strategies for effective human-computer interaction* (5th) (M. Hirsch, Ed.). Pearson Addison-Wesley. Retrieved from http://wps.aw.com/aw_shneiderman_dtui_5/

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, *28*(4), 409–445. doi:10.1162/089120102762671936

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *EMNLP '06: proc. 2006 conference on empirical methods in natural language processing* (pp. 103–110). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1610075.1610091

Thomson Reuters. (2011a). EndNote [Software]. Retrieved from http://www.endnote.com

Thomson Reuters. (2011b). ISI Web of Knowledge. Retrieved from http://isiwebofknowledge.com

Zajic, D., Dorr, B., & Schwartz, R. (2004). BBN/UMD at DUC-2004: Topiary. In *HLT/NAACL-DUC '04: proc. HLT/NAACL 2004 workshop on document understanding* (pp. 112–119). Boston, MA. Retrieved from http://www.umiacs.umd.edu/~bonnie/publications.html

Zajic, D., Dorr, B. J., Lin, J., & Schwartz, R. (2007). Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management*, *43*(6), 1549–1570. doi:10.1016/j.ipm.2007.01.016

Zajic, D. M., Dorr, B. J., Schwartz, R., Monz, C., & Lin, J. (2005). A sentence-trimming approach to multi-document summarization. In *HLT/EMNLP '05: proc. HLT/EMNLP 2005 workshop on text summarization* (pp. 151–158). Vancouver, Canada. Retrieved from http://www.casl.umd.edu/node/719