Evaluating Visual and Statistical Exploration of Scientific Literature Networks

Robert Gove[1,2,4], Cody Dunne[1,2,4], Ben Shneiderman[1,2,4], Judith Klavans[2,3], and Bonnie Dorr[2,3,4]
[1]Human-Computer Interaction Lab, [2]Institute for Advanced Computer Studies
[3]Computational Linguistics and Information Processing Lab [4]Department of Computer Science
University of Maryland, College Park, MD, USA
{rpgove, cdunne, ben}@cs.umd.edu, jklavans@umd.edu, bonnie@cs.umd.edu

***Abstract*—Action Science Explorer (ASE) is a tool designed to support users in rapidly generating readily consumable summaries of academic literature. It uses citation network visualization, ranking and filtering papers by network statistics, and automatic clustering and summarization techniques. We describe how early formative evaluations of ASE led to a mature system evaluation, consisting of an in-depth empirical evaluation with four domain experts. The evaluation tasks were of two types: predefined tasks to test system performance in common scenarios, and user-defined tasks to test the system's usefulness for custom exploration goals. The primary contribution of this paper is a validation of the ASE design and recommendations to provide: easy-to-understand metrics for ranking and filtering documents, user control over which document sets to explore, and overviews of the document set in coordinated views along with details-on-demand of specific papers. We contribute a taxonomy of features for literature search and exploration tools and describe exploration goals identified by our participants.**

***Keywords*-Empirical evaluation; graphical user interfaces; information visualization; literature exploration**

I. Introduction

A common challenge for researchers and scholars is to quickly learn about a new research area. A researcher who wishes to apply a state-of-the-art technique to his research will need to search for leading papers and recent breakthroughs. Review panel members who are reviewing grant proposals from unfamiliar fields may need to identify open questions, current trends, and emerging fields. Graduate students who need to become familiar with research in their chosen areas may search for historical papers, leading authors, appropriate publication venues, and current research methodology.

These tasks can be challenging due to disparate sources of literature and the relative lack of powerful tools for exploring and summarizing academic literature. Users with different profiles may explore academic literature to answer different but overlapping questions. Therefore, tools for exploring literature should ideally support a variety of tasks and levels of users, which makes them more challenging to design.

Digital libraries continue to expand due to new literature being written and old literature continuously being digitized. Although improving the accessibility of literature should benefit researchers, it puts pressure on literature exploration tools to support larger datasets, thus making in-depth exploration of digital libraries a more challenging process.

Existing digital libraries and reference managers support a wide variety of features, with surprisingly little overlap, possibly indicating no consensus about what features are most salient for exploring academic literature. Section II describes current systems for searching and summarizing academic literature. Search engine-based tools naturally center around searching, but users may not have a clear goal in mind, or the goal may not be expressed in a query that a search engine can easily answer, such as identifying the most prominent authors or research communities in a given field.

Action Science Explorer (ASE) [1], [2] is a new system that integrates many capabilities from other literature-centric information retrieval systems. These include a citation network visualization, searching, citation sentences, and a reference manager. In addition, it provides several unique capabilities among literature search and exploration tools, such as ranking and filtering papers by citation network statistics, automatically detecting and visualizing clusters of papers, and creating computer-generated summaries of the clusters (further discussed in Section III).

This paper's primary contribution is validation of the choices made in ASE's design and design recommendations for literature exploration systems, discussed in Sections IV and V. Especially useful features for information seekers are the visual ranking and filtering of papers by citation network statistics and the control over what document sets to explore. We contribute a taxonomy of features for literature search and exploration tools and describe exploration goals identified by our participants.

II. Related Work

According to Hearst [3], information seekers (ISers) typically search by recognizing an information need, specifying the query, evaluating the results, and possibly reformulating the query. However, when users are unsure what they are looking for, or how to describe it, they may prefer to participate in exploration or browsing behavior instead. Pirolli and Card [4] discuss information foraging, where analysts select documents with a potentially high profit and narrow the set down before finally reading the refined set of documents. In this way analysts follow a "scent" or cue directing the search process. Bates [5] presents several information search stratagems and gives examples, such as the Area Scan where users identify a general area of interest and then browse in that area.

A. Academic Literature Exploration

Several tools exist for exploring and summarizing academic literature, each with different strengths and features. This list is a sample of the available tools and features.

CiteSeerX is a digital library for scientific literature [6], [7] that incorporates citation statistics, lists of citing papers, and the citation context (i.e. the excerpt of the citing paper that surrounds the citation). These summaries and statistics can reveal a paper's impact and intellectual contribution. Google Scholar [8], Microsoft Academic Search (MAS) [9], and ArnetMiner

[10] are similar systems. Google Scholar applies the Google search interface to articles, patents, and court cases, but with a smaller variety of statistics and metadata than CiteSeerX. MAS shows paper data similar to CiteSeerX, but also provides author statistics such as publication count, total citations, and number of co-authors. Similarly, ArnetMiner [10] includes paper and author statistics, but also models academic social networks and supports expertise and association search. These capabilities are potentially highly valuable for exploring academic literature, but in CiteSeerX, MAS, and ArnetMiner they are spread across different pages—each view is essentially limited to one entity rather than an entire discipline. Thus, CiteSeerX, Google Scholar, MAS, and ArnetMiner can be useful tools for summarizing individual papers or authors, but may not be ideal for summarizing an entire corpus.

GoPubMed [11] shows publications by year and country (both in numeric form and non-interactive visualizations) and by journal and keywords (numeric form only), plus a network showing coauthorship of top authors. However, users must reload the page to update the visualization, which slows down user interaction. Furthermore, these data are aggregates across authors, journals, and search terms rather than per paper, which potentially limits the analysis capabilities.

Web of Knowledge [12] can create a hyperbolic citation tree that shows everything a paper cites and everything that cites the papers. These citation trees support a limited amount of dynamic interaction; however, some interactions, such as filtering by publication year, require recreating the citation tree that slows down the interaction and exploration process. Web of Knowledge also offers other capabilities that can be found in other software, such as ACM Portal and IEEE Xplore. These capabilities include a list of documents that cite a paper c, some document statistics, and ranking capabilities. OntoDoc [13] is a different kind of system for querying digital libraries, which can query for documents with certain visual content. E.g., black color, two columns of English text, and an image.

One of the limitations of the above systems is that there is no way to load additional documents, or to filter out all but a subset of the documents. The tools described below allow users to define custom libraries.

Reference managers such as JabRef [14], Zotero [15], and Mendeley [16] also provide search features and some limited summarization capabilities. For example, users can import a corpus into Mendeley and then search the full-text of documents, write notes on a document, and view limited statistics such as the number of publications per author. However, many of these features in reference managers are no better for summarizing groups of papers than digital libraries, and in some ways are less powerful because reference managers do not tend to provide lists of citing papers or citation context.

Recommender systems can also be used to aid the exploration of academic literature [17]–[19]. These systems recommend papers based on a given input, but such systems do not necessarily give users an overview of a domain, nor do they freely allow users to explore the literature.

Boyack and Börner [20] used VxInsight [21] to visualize interconnections between research areas, institutions, grants, publications, and authors. The relationships were visualized as 3D landscape maps with points representing keywords, institutions, and other fields. Boyack and Börner's primary goal was to explore these relationships. Although their visualizations yielded interesting results, a purely visualization-driven interface would have difficulty revealing details of the research results in a given domain.

TABLE I
SYSTEM CAPABILITIES, GROUPED BY SYSTEM TYPE (SEARCH ENGINES, REFERENCE MANAGERS, AND SUMMARIZERS AND RECOMMENDERS).

	Search Engines						Ref. Mangrs				Sum. & Rec.			
Functionality	CiteSeerX	Google Scholar	GoPubMed	Web of Knowledge	ACM Portal	IEEE Xplore	JabRef	Zotero	Mendeley	EndNote	XploreMed	Recommender Sys.	NewsInEssence	Trimmer
Summary of textual excerpts													●	●
Custom database							●	●	●	●	●	●	●	●
Create notes							●	●	●	●				
Ranking	●			●	●	●	●	●	●	●				
Full-text search	●	●						●	●	●			●	
Document recommendations	●	●	●	●								●	●	
Document statistics	●	●	●	●	●	●								
"Cited by" list	●	●		●	●	●								
Search excerpts	●	●							●					
Corpus statistics			●						●		●			
Keyword summary			●						●		●			
Citation visualization			●	●										
Citation context	●													

Tools for summarizing literature include XplorMed, which summarizes search results by keywords and word context [22]. Users supply their own database and perform analyses on an abstract. Results show the abstract, keywords, keyword chains, and keyword frequencies in the abstract, but no visualizations of the data. NewsInEssence [23] summarizes news topics, but might not reveal the evolution of a field because it does not necessarily account for the overall structure of the source text. NewsCube [24] allows aspect-level news browsing and classifies articles by similarities of aspects, or viewpoints, that allows partitioning by topic and by author perspective. However, neither summarization nor aspect-level browsing alone will provide both overviews and details-on-demand of a pool of articles. The iOPENER project explored citation network visualizations to reveal research fronts and their relationships on a semantic substrate that grouped related papers in separate screen regions [25].

B. Tool Taxonomy

Table I summarizes the above systems by the following capabilities, defined in terms of a given paper c, a user-specified corpus C (that includes c), and a superset corpus S (e.g., the Association of Computational Linguistics Anthology):

"Cited by" list: A document list in C or S that cite c.

Citation context: The citing text that shows the context of the citation of c.

Search excerpts: Query results that show text excerpts highlighting the query keywords in c.

Citation visualization: A visualization showing a partial or complete citation network.

Custom database: The capability to load custom document database S into the system.

Full-text search: Query tool searching text in c as well as the metadata associated with c.

Create notes: Tool for creating notes in the system.

Corpus statistics: Statistical information about C or S.

Document statistics: Statistical information about c.

Keyword summary: List of the most prominent keywords or phrases in c, C, or their metadata.

Document recommendations: Recommendations of documents in C or S of likely interest to users.

Ranking: A document ranking in C or S, using features such as publication year, author, citation count, etc.

Summary of textual excerpts: Automatic summaries of textual excerpts from c or documents citing c.

Systems such as CiteSeerX tend to support literature exploration and overviews of articles and authors, but do not usually support summarizing entire fields or corpora using network statistics or clustering algorithms. Reference managers support custom-built databases and sorting references by various properties, but generally lack information on citations between papers. Recommender and summarization systems provide automatic summaries of corpora, but the ones surveyed here do not support citation-level summaries or network statistics. Overall, each class of system tends to support only a specific subset of capabilities.

Most of these applications lack visualizations and the textual context for individual citations. Each application has a different target audience, but there seems to be a lack of support for effectively summarizing and exploring academic literature for a given domain.

C. Evaluation Strategies

Evaluating information visualizations is challenging: widely used quantitative techniques in tightly controlled laboratory conditions may not be adequate for capturing important aspects of tool usage, such as creativity and discovery [26]–[28]. Common metrics in HCI research often focus on measuring the time to complete a task or the number of errors and then comparing the results to a benchmark [26]. One alternative is to quantitatively or qualitatively analyze the insights gained from the visualization [29]–[32].

Qualitative evaluation methods for information visualization tools are becoming more common [28], which include longitudinal studies [32]–[34], Grounded Theory [35], and case studies [33], [34], [36]. This reflects a growing shift that emphasizes the need for data analysts to use tools under naturalistic conditions similar to their everyday situations [28], [37]. Quantitative methods such as surveys and questionnaires can also be used to augment qualitative methods [38].

Several researchers stress the need for research participants analyze their own data, which keeps participants motivated and gives the tool significance [28], [32], [35], [37]. Users may be less likely to make discoveries if the discoveries and data are less significant to the users.

III. A New Approach: Action Science Explorer

Action Science Explorer[1] (ASE) [1], [2] is a tool that can be used to explore and create readily-consumable surveys of scientific literature domains. The tool integrates citation network link mining and visualization, citation sentence extraction, ranking and filtering by network statistics, automatic document clustering and summarization, and reference management. This is a combination of several features seen in other literature exploration and search tools with additional novel capabilities. A screenshot is shown in Figure 1.

ASE makes use of the JabRef reference manager [14] which allows for searching, grouping papers, sorting the bibliography, viewing abstracts, creating text annotations, and viewing the full-text PDF. Because JabRef is built around the BibTeX bibliography format, JabRef can easily be used in conjunction with Microsoft Office, LaTeX, and OpenOffice.org with the appropriate plug-ins.

ASE also integrates SocialAction [39], a powerful network analysis tool that provides a node-link visualization of the citation network, paper ranking and filtering on network statistics, automatic cluster detection and display, and scatter plots of network measures and paper attributes. These tools provide an overview of individual research communities in the corpus and a context for individual documents.

A third set of views provides several text views and summaries of the documents. If a paper or cluster of papers is selected, the In-Cite Text view shows a list of all sentences from other papers that are citing the selected paper(s). Each of these sentences is in turn a hyperlink. Clicking on one will open the full-text of the source paper in the Out-Cite Text view, which includes citation highlighting to show other hyperlinked citations to other papers. If users select a cluster of documents, then the In-Cite Summary view shows an automatically generated summary of all the sentences shown in the In-Cite Text view.

All of these views are linked together, showing multiple coordinated views of the corpus. When a paper is selected in the reference manager, it becomes highlighted in the citation network visualization. If users click on a node in the citation network, the citation text is shown in the In-Cite Text view. If users click on a link in the Out-Cite Text view, the corresponding paper is selected in the network visualization. Each view can be hidden or rearranged, allowing users to tailor their workspace to the screen space available and their current focus.

Below are detailed descriptions of some of the most prominent features of ASE, again defined in terms of c, C, and S:

Attribute Ranking: Ranks and provides dynamic filtering of the nodes (papers) by network statistics, such as document set in-degree (number of times c is cited in C) and out-degree (number of papers in C cited by c), corpus in-degree (number of times c is cited in S) and out-degree (number of papers in S cited by c), as well as betweenness centrality, clustering coefficient, and the like.

Find Communities: Clusters papers using Newman's community detection algorithm [40] to cluster documents in C and shows the clusters using convex hulls in the network visualization. A user-movable slider sets clustering cutoffs, ranging from a single large loosely connected cluster to a single small tightly connected cluster (with several mid-sized clusters halfway in between).

Search: Can search any field in the bibliographic entry for c, including title, author, abstract, keywords, and user-added notes, using regular expressions.

Sorting: Sorts bibliographic entries in C by any field, such as title or author.

Groups: Users can manually create hierarchical groupings of papers in C.

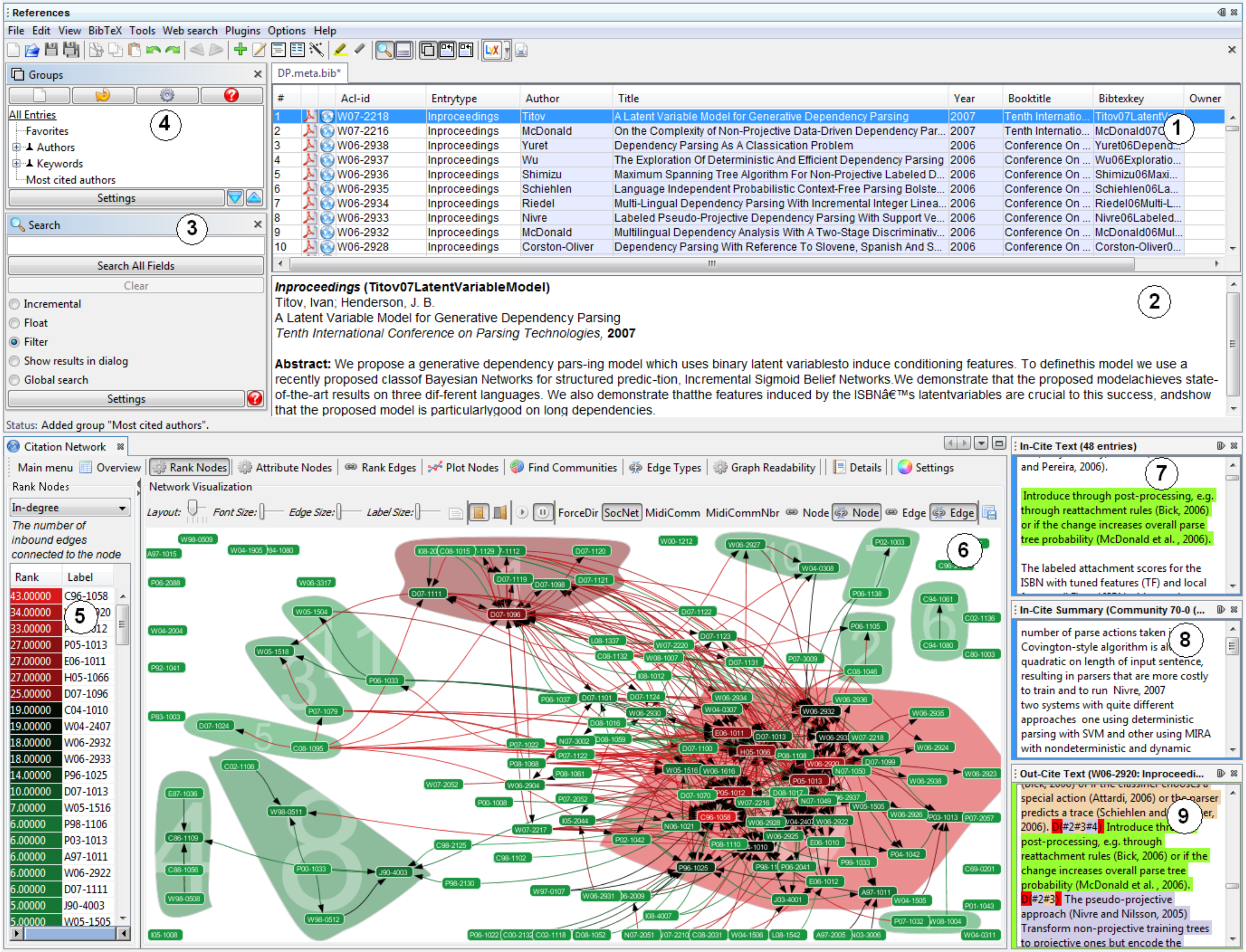
In-Cite Text: Shows sentences from other papers in C or S that cite c. Each sentence is a hyperlink that opens the full-text of the citing paper in the Out-Cite Text view and highlights the citation.

In-Cite Summary: Displays an automatically generated summary of all sentences that cite c, using Multi-Document Trimmer [41].

Out-Cite Text: Provides the full-text of the source paper c, with sentence highlighting to show other hyperlinked sentences to other papers C.

These features provide a multi-faceted approach to literature exploration and analysis. Attribute Ranking, Filtering,

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



#	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	20
33.00000	P 12
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

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

number of parse actions taken 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

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

Fig. 1. The Action Science Explorer system loaded with the 147-paper dependency parsing data set: (1) the bibliography list, (2) bibliographic entry details, (3) bibliographic search, (4) bibliographic groups, (5) the attribute ranking and filtering view, (6) the citation network visualization, (7) the In-Cite Text, (8) the multi-document In-Cite Summary of a cluster, and (9) the full-text article with linked citations.

and Finding Communities potentially allow users to quickly identify papers with key metrics and groups of papers clustered around a common theme.

IV. Evaluation

Early formative evaluations consisted of four participants over a period of about one year: two participants used Action Science Explorer (ASE) in multiple 1–2 hour sessions with assigned tasks, and two other participants used ASE together to explore a dataset freely and report on their findings. These prior evaluations helped guide the development of ASE and provided a foundation for the evaluation presented in this paper.

To evaluate the usability and effectiveness of the improved (ASE) after the prior evaluations, we conducted an insight-based study. Evaluation sessions included two predefined tasks for all participants to determine how some task performance generalizes across users, and user-defined tasks were used to assess ASE's ability to help users make discoveries relevant to themselves.

Participants: The qualitative evaluation sessions were performed with two male and two female participants P1, P2, P3, and P4. Two were current Computer Science PhD students and the others recent graduates. Ages ranged from 24–32 years old.

All of the participants are currently working in computational linguistics. P2 and P4 already had some experience with dependency parsing, describing themselves as having "a little bit" of experience, and being "quite familiar" respectively. P1 and P3 had no prior experience with dependency parsing. P2 also had some previous experience using JabRef, but did not appear to have experience with anything except the basic features for adding references and editing fields.

Experimental Design: Evaluations were conducted on an Intel Core i3 2.26 Ghz laptop with 4 GB of RAM. ASE was displayed on an external 30-inch monitor at a resolution of 1920x1080. The facilitator limited each session to two hours, and began with a 30 minute training session. Participants were asked to perform some predefined tasks, which usually lasted for one hour. For the last 30 minutes, participants were asked to perform tasks they thought would be useful. Participants were asked to think aloud while they performed the tasks, while the observer took notes about the participants' thoughts and actions. Finally, participants commented on their experience.

The evaluation focused on 147 papers from the Association for Computational Linguistics (ACL) Anthology Network [42] that were isolated by a keyword search for "dependency parsing". The data includes fields such as author, title, publication year, publication venue, and a field for the unique ACL ID. The ACL ID is used for the network visualization label, and the ID is based on the publication venue and year (e.g. W03-0501 is taken from the HLT-NAACL 03 Text Summarization Workshop in 2003).

Procedure: Participants were recruited at a computational linguistics seminar, and $30 USD was offered. The participants signed informed consent forms and were given the opportunity to withdraw from the study at any time, per IRB requirements.

The reason for recruiting participants with experience in computational linguistics was to address the motivation issue discussed in Section II-C. Participants familiar with computational linguistics will be more likely to understand and want to explore documents on dependency parsing.

Each participant was run in a separate evaluation session. During the training stage, the participants were shown video clips describing ASE and demonstrating its features. After each video, the participant practiced by performing the tasks shown in the videos. The participants were also given opportunities to ask questions if they did not understand the tool or its features.

The participants were asked to perform the following predefined tasks: (1) Identify and make note of the important authors and papers, and (2) find an important paper and collect evidence to determine why it is important. This provided a common benchmark for performance of basic tool functionality, as well as showing variation in the depth of analysis that participants were able to achieve.

Finally, the participants were asked what they would want to find out about the dataset. One or more of these goals were selected by the facilitator as tasks for that individual participant, and the participant was asked to perform that task or tasks. This increased the relevance to each participant in their evaluation session.

V. Results

There were two task sets during the evaluation sessions: predefined and user-defined.

Predefined tasks: During the first predefined task (to identify prominent papers and authors), all participants created two groups in the reference manager for storing the most important authors and papers. P1 and P3 relied mostly on attribute ranking and filtering to determine prominence, whereas P2 and P4 relied more on their prior knowledge of the field.

P1 took the simplest approach and initially only used attribute ranking to select prominent papers and authors who wrote highly cited papers. Later he gained deeper insight by scanning the In-Cite Text, observing authors who were frequently cited by other papers. He also discovered an author who seemed to have done a lot of work on efficient algorithms, and indicated the author must be important, regardless of citation count.

In order to identify prominent authors, P3 also viewed rankings on a per-author basis to identify the most frequently-cited authors. She described a method of using Find Communities and looking in each cluster for highly cited authors, but had trouble completing the task.

Initially, P2 only used the reference manager to view the list of authors and select ones she already knew were key. She tried to augment her approach by finding all papers of a certain author in the reference manager and selecting them to view their in-degree ranking in the ranked list. However, due to a software bug, her selection was not reflected in the attribute ranking, which made it difficult to see the ranking for each authors' papers though the nodes were still colored by their ranking. To identify important papers, P2 wanted to rank papers both by in-degree and publication year to see what years had the most highly cited papers.

P4 had a similar approach to P2, but only used JabRef to sort by author name and visually scan the list to select authors and papers that he already regarded as important. He did not use any other part of the tool to verify his notions of importance, indicating that he was confident in his selections and did not expect to make any additional discoveries.

For the second predefined task (finding evidence to determine why a paper is important), P2 and P4 continued their pattern of relying on their prior knowledge. P4 did not use the tool at all and described from memory why his choice was important. P2 combined her existing knowledge with attribute ranking by corpus in-degree, and was able to disregard some well-ranked papers because she felt those papers are usually

TABLE II

CAPABILITIES USED BY EACH PARTICIPANT FOR THE TWO PREDEFINED TASKS ((1) IDENTIFYING PROMINENT AUTHORS AND PAPERS, AND (2) COLLECTING EVIDENCE). **AR** = ATTRIBUTE RANKING, **FC** = FIND COMMUNITIES, **SEARCH** = REFERENCE MANAGER SEARCH, **SORT** = REFERENCE MANAGER SORTING, **GRP** = REFERENCE MANAGER GROUPS, **ICT** = IN-CITE TEXT, **ICS** = IN-CITE SUMMARY

	First Task							Second Task						
Participant	AR	FC	Search	Sort	GRP	ICT	ICS	AR	FC	Search	Sort	GRP	ICT	ICS
P1	●				●	●			●			●		
P2	●		●		●			●						
P3	●							●						
P4				●										

cited out of obligation rather than relevance. She ultimately chose a paper and justified her claim by filtering the attribute ranking to show only the highest corpus in-degree and then seeing in the visualization that this paper linked two of the most highly cited papers in this filtered subset.

P3 used a similar approach as P2, and ranked nodes by in-degree. Then she selected a highly cited paper and justified her claim by pointing out in the visualization that it is cited by another highly cited paper.

P1 changed his approach several times while performing the task. He began by enabling Find Communities and viewing the In-Cite Summary on the largest community, but did not find anything that he thought was helpful. Then he said he felt that the most important paper would be the most recent paper that gave an efficient algorithm for dependency parsing. He wanted to find a list of all papers that cited a given paper and then sort the results by publication year, but he had difficulty with this because there is no automated way to do it. He then went back to Find Communities and began searching through the largest community by looking at the incoming citation sentences. He eventually chose C96-1058 and H05-1066 based on the sentences shown in the In-Cite Text view that talk about efficient algorithms, but he could not satisfy his original search for a most recent paper on efficient algorithms.

User-defined tasks: Users were asked what kinds of questions they would want to ask while exploring this dataset. Below is a summary of these questions:

- Identify the historical foundations of the field.
- Identify big breakthroughs, and the state of the art.
- Find relationships between communities and clusters, such as papers that use the same research methods.
- Discover frequent co-authors.
- Find review papers.
- Search for specific aspects (e.g. models or applications).
- Identify different research methods used in the literature.
- Identify student-advisor relationships in papers.

The tasks provided by each participant were unique, aside from P1 and P4 both wanting to identify relationships among papers within communities. However, some participants identified similar questions, such as discovering frequent co-authors and discovering student-advisor relationships.

To test whether users are able to answer their own questions using the tool, the facilitator chose some of the tasks identified by each participant and asked that participant to perform the task(s). P1 did not perform his task because the allotted time for the session had expired by this point.

P2's task was to try to search for papers she would start reading to identify the state of the art in dependency parsing, since this was one of the tasks she had identified. She first tried ranking papers by publication year and trying to look at the In-Cite Text for the most recent papers, but they did not have In-Cite Text because they had no citations. She then continued this approach without using the Attribute Ranking by selecting papers in the visualization and examining their publication year and In-Cite Text. She did not appear to have any specific heuristic for selecting papers, but eventually chose a paper from 2007. She commented that finding the state of the art can be difficult if it is too new because it will not yet be cited by many papers, which highlights a general problem in identifying emerging research fronts. She created a group to store her results, and examined the paper titles and abstracts for relevance, and eventually opened the full PDF version of candidate papers to scan them. In the end, she successfully identified papers that she wanted to investigate later.

P3 was assigned two additional tasks. First, she was asked to determine if the most important author was still active. To accomplish this, she wished to rank papers both by in-cites and by publication year. This action is not directly supported in the tool, so after a few attempts to do this, she changed her approach because she felt that recent publications might not be highly cited. She then sorted the papers by author and looked for authors with many publications who also have a recent publication, allowing her to identify two authors.

P3's second task was to find review papers. Her approach was to rank nodes by corpus out-degree because a review paper will cite lots of other papers. After looking at the top papers and not finding a review paper, she also used the search feature to search for the keywords "tutorial" and "survey" but did not find any matches. After discussing the dataset with someone knowledgeable of dependency parsing after the evaluation session, it is likely that there is no survey paper on dependency parsing in this dataset.

P4's task was to find any unifying themes in any of the communities. He used the Find Communities feature and noticed that most of the papers in the biggest community cited the same central papers. He proceeded to look at the titles of the papers in the community, and he then mentioned that it was hard to find a unifying theme by looking at the paper titles. Eventually he said "It doesn't jump out at me why it clustered these things [this way]." At this point he felt that he had looked as much as he could, and he gave up.

Additional tasks: Due to P4's familiarity with dependency parsing, he performed four additional tasks. First, determine if any interesting research used any of the most important papers. He began by selecting a paper, P05-1013, in the visualization and looking at recent citations: he would examine the In-Cite Text, click on the hyperlink, and then look at the publication at the top of the Out-Cite Text. He expressed frustration that if he opened the PDF for any of these citing papers that he had to remove focus from P05-1013. After using the procedure for a few papers and opening a few PDFs to see if authors were citing P05-1013 in a way indicating something interesting, he isolated D07-1123 as an interesting advance of P05-1013.

P4's second task was to use the tool to support his claims about the most important authors from the first pre-defined task. He first ranked the nodes by document set in-degree and corpus in-degree, and mostly he used a single paper from each author as a representative for the author's body of work; however, he expressed that he wanted to group papers by author and look at rankings just for a specific author's papers. By looking only at the document set in-degree and corpus in-degree rankings, he said that he believes his choice of Schank is not supported; he felt the paper is classic, but it has a low document set in-degree and corpus in-degree ranking. For another author, he felt that the paper was actually ranked higher than it should have been, but nonetheless his selection was justified. For the other three papers—P05-1013, C96-1058, and P96-1025—he felt that the attribute rankings confirmed his choices, and that overall his choice of authors was confirmed.

For the third task, P4 needed to use the tool to support his claims about important papers. This time he used only attribute ranking by corpus in-degree, and felt that eight of his eleven choices were confirmed by the attribute rankings.

P4's final task was to use ASE to summarize the literature. He first commented that ranking the nodes and finding the top paper could be a good place to begin. However, he decided to use the Find Communities feature and look at the In-Cite Summary for the largest community. He noticed several things in the automatically generated summary that were factually incorrect, and decided that the summary was not helpful. At this point the allotted session time expired and the participant did not have time to complete the task.

VI. Discussion

Overall, participants were able to quickly grasp the basics of ASE's reference manager and network visualization components. Some participants immediately began using somewhat more advanced features, e.g., P1 used searches almost from the beginning. However, participants tended to use the same functionality at the beginning of the sessions as at the end.

By examining the functionality used per task shown in Table II it is clear that by far the most used feature was attribute ranking. Participants almost uniformly used it to determine author and paper importance. Even when tasks did not directly involve finding an important paper participants still relied heavily on attribute ranking, such as when searching for review papers. This may be partly because summaries of literature often focus on the most salient aspects of the corpus and users may believe those papers will be cited most often.

As Hearst discusses [3], there are theories describing how ISers navigate based on information "scent." For example, P2 used the publication year and In-Cite Text to build a set of papers that she then explored in more depth by examining the full PDF version to identify the state of the art in dependency parsing. It is possible that participants were so eager to use the Attribute Ranking feature because it provided very powerful quantitative clues to begin their navigation process and it offered ISers a simple way to direct their search during the Exploration stage of Kuhlthau's [43] description of information seeking behavior.

It could also be that Attribute Ranking offered a fast easy-to-use way for participants to filter out potentially unimportant papers and allow the participants to focus on exploring portions of the literature that could be more important. This is similar to the Area Scan stratagem described by Bates [5] where users browse in a general area after identifying an area of interest.

In many cases, participants did not seem to use multiple features in the tool to drill down and find specific answers to questions; they tended to use a single tool feature, such as Attribute Ranking, to support their claims. For example, when participants selected an important paper most of the participants only used Attribute Ranking by document set in-degree to justify their claim. It is possible that participants found Attribute Ranking to be very accessible because it is simple to use and easy to interpret.

The network visualization overview was also easily understood by the participants, and they frequently used it to orient themselves in the data set. Participants even used the visualization to support claims about the data set, e.g., P3 used the visualization to show that she had selected a highly cited paper that was cited by another highly cited paper.

Although the Out-Cite Text view is potentially useful, it was sometimes problematic. For example, participants had difficulty understanding what the Out-Cite Text was and which paper it came from. Furthermore, when using the Out-Cite Text view to change focus in the network diagram to a different paper, users had difficulty returning focus to the previously selected paper because there was no convenient way to undo the change of focus. The task might have been made easier with a better indication of relationships between views, an ability to create a new corpus from the first selected paper and its citing papers, and an undo/redo functionality.

Many of the analyses performed by the participants may not have been possible, or would have been much more difficult, if the tool did not allow users to use a custom database of documents. Even so, participants frequently expressed interest in analyzing only a subset of the papers in the network. Frequently participants wanted to do this to show a given node and all of its neighbors. This underscores the importance of allowing users to select a custom corpus of documents and drill down to perform fine-grained analysis.

The In-Cite Summary is a feature that participants wanted to use. During the evaluation the output quality of the summarization algorithm was poor. Trimmer was originally designed for news articles, not citation sentences. This is an avenue of on-going research, and we have already achieved much better results after tailoring Trimmer to this task.

Three of the four participants said that they liked the tool and thought it was useful, in particular for searching for important papers. However, P4 was skeptical about the tool's utility: he felt that the interface is too "busy", and that the tool might not be useful unless he had a goal in mind and he already knew a few papers as a starting point. P2 also commented that the interface is complicated, but she had an overall positive response, enjoyed using the network visualization, and was very enthusiastic about the tool. This difficulty may be reduced through a guided analysis, where new views are opened up after users are comfortable with a base set.

A limitation of the user study was the size of the corpus: 147 papers. This demonstrates a particular use case for visualizing the results of a search in a larger corpus, but may not be indicative of exploring the entire corpus. However, the small corpus did not appear to affect understanding. A further limitation was the small number of participants, but their domain knowledge and willingness to work on complex tasks provided us with a wealth of insights about ASE's strengths and weaknesses.

In particular, we found that Attribute Ranking and filtering were very popular and provided users with a simple method to begin their searches, and it became clear that users wish to specify, at a fine-grained level, which documents to explore and analyze. This study also provided clear guidance for our work and specific recommendations to designers of literature exploration systems:

- Allow users control over what documents to explore, both to select an initial document set as well as drill down on a subset of documents.
- Provide easy-to-understand metrics for ranking documents.
- Give users a way to see overviews of the corpus using multiple coordinated views.

- Incorporate undo/redo functionality to navigate the history.
- Include capabilities to create groups of documents and annotations for documents.

VII. Conclusion

Between the early formative evaluation and the mature evaluation, we developed a strong understanding of the strengths and weaknesses of ASE. These evaluations identified a list of questions that ISers ask when exploring new scientific literature domains. Furthermore, it found that ASE offers novel and useful features to ISers, especially visual paper ranking and filtering. We also suggest that designers provide support for users to select a subset of the citation network and perform analyses only on that subset of documents. We intend to perform another user evaluation of the system to analyze the potential benefits of this functionality. Our taxonomy of literature search and exploration tools based on their capabilities should aid users in future tool comparisons, including a planned comparative evaluation of ASE with existing tools.

Acknowledgment

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 Univ. of Michigan and Univ. of Maryland as IIS 0705832. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr, "Rapid understanding of scientific paper collections: integrating statistics, text analysis, and visualization," University of Maryland, Human-Computer Interaction Lab Tech Report HCIL-2011, 2011.

[2] C. Dunne, B. Shneiderman, B. Dorr, and J. Klavans, "iOpener Workbench: Tools for rapid understanding of scientific literature," in *Human-Computer Interaction Lab 27th Annual Symposium*, University of Maryland, College Park, MD, 2010. [Online]. Available: ftp://ftp.umiacs.umd.edu/pub/bonnie/iOPENER-5-27-2010.pdf

[3] M. A. Hearst, *Search User Interfaces*. Cambridge University Press, 2009. [Online]. Available: http://searchuserinterfaces.com/book/

[4] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proc. Int. Conf. Intelligence Anal.*, 2005.

[5] M. J. Bates, "Where should the person stop and the information search interface start?" *Inf. Process. Manage.*, vol. 26, pp. 575–591, 1990.

[6] K. D. Bollacker, S. Lawrence, and C. L. Giles, "CiteSeer: an autonomous Web agent for automatic retrieval and identification of interesting publications," in *Proc. int. conf. Autonomous Agents*, 1998, pp. 116–123.

[7] C. L. Giles, K. D. Bollacker, and S. Lawrence, "CiteSeer: an automatic citation indexing system," in *Proc. ACM conf. Digital Libraries*, 1998, pp. 89–98.

[8] Google, "Google Scholar," http://scholar.google.com/, Mar, 2011.

[9] Microsoft, "Microsoft Academic Search," http://academic.research.microsoft.com/, June 2011.

[10] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: extraction and mining of academic social networks," in *Knowledge Discovery and Data Mining*, 2008, pp. 990–998.

[11] Transinsight, "GoPubMed," http://www.gopubmed.org/, Mar, 2011.

[12] Thomson Reuters, "ISI web of knowledge," http://www.isiwebofknowledge.com/, Mar, 2011.

[13] A. Malizia, "A cognition-based approach for querying personal digital libraries," in *Visual Languages and Human Centric Computing*, 2004, pp. 283–284.

[14] JabRef Development Team, *JabRef*, 2011. [Online]. Available: http://jabref.sourceforge.net

[15] Center for History and New Media, "Zotero," http://www.zotero.org/, Mar, 2011.

[16] Mendeley Ltd, "Mendeley," http://www.mendeley.com/, Mar, 2011.

[17] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl, "Enhancing digital libraries with TechLens+," in *Proc. Joint Conf. Digital libraries*, 2004, pp. 228–236.

[18] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl, "On the recommending of citations for research papers," in *Proc. conf. Comp. supported cooperative work*, 2002, pp. 116–125.

[19] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *Proc. int. conf. World wide web*, 2010, pp. 421–430.

[20] K. W. Boyack and K. Börner, "Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers," *JASIST*, vol. 54, pp. 447–461, 2003.

[21] G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N. Wylie, "Knowledge mining with vxinsight: Discovery through interaction," *J. Intelligent Info Sys*, vol. 11, pp. 259–285, 1998.

[22] C. Perez-Iratxeta, A. J. Perez, P. Bork, and M. A. Andrade, "Update on XplorMed: a web server for exploring scientific literature," *Nucleic Acids Research*, vol. 31, pp. 3866–3868, 2003.

[23] D. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn, "Newsinessence: summarizing online news topics," *Commun. ACM*, vol. 48, pp. 95–98, Oct 2005.

[24] S. Park, S. Lee, and J. Song, "Aspect-level news browsing: understanding news events from multiple viewpoints," in *Proc. int. conf. Intelligent user interfaces*, 2010, pp. 41–50.

[25] A. Aris, B. Shneiderman, V. Qazvinian, and D. Radev, "Visual overviews for discovering key papers and influences across research fronts," *JASIST*, vol. 60, pp. 2219–2228, Nov 2009.

[26] E. Bertini, A. Perer, C. Plaisant, and G. Santucci, "BELIV '08: Beyond time and errors: novel evaluation methods for information visualization," in *CHI '08 extended abstracts on Human factors in computing systems*, 2008, pp. 3913–3916.

[27] C. Chen and M. P. Czerwinski, "Empirical evaluation of information visualizations: an introduction," *Int. J. Hum.-Comput. Stud.*, vol. 53, no. 5, pp. 631–635, 2000.

[28] B. Shneiderman and C. Plaisant, "Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies," in *Proc. 2006 AVI workshop on BEyond time and errors*, 2006, pp. 1–7.

[29] P. Saraiya, C. North, and K. Duca, "Comparing benchmark task and insight evaluation methods on timeseries graph visualizations," in *Proc. 2010 AVI workshop on BEyond time and errors*, 2010.

[30] J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko, "Understanding and characterizing insights: how do people gain insights using information visualization?" in *Proc. Conf. BEyond time and errors*, 2008, pp. 1–6.

[31] L. Shupp, C. Andrews, M. Dickey-Kurdziolek, B. Yost, and C. North, "Shaping the display of the future: The effects of display size and curvature on user performance and insights," *Human-Computer Interaction*, vol. 24, pp. 230–272, 2009.

[32] P. Saraiya, C. North, V. Lam, and K. Duca, "An insight-based longitudinal study of visual analytics," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, pp. 1511–1522, 2006.

[33] J. Seo and B. Shneiderman, "Knowledge discovery in high-dimensional data: case studies and a user survey for the rank-by-feature framework," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, pp. 311–322, 2006.

[34] A. Perer and B. Shneiderman, "Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis," in *Proc. SIGCHI conf. Human factors in computing syst.*, 2008, pp. 265–274.

[35] S. Faisal, B. Craft, P. Cairns, and A. Blandford, "Internalization, qualitative methods, and evaluation," in *Proc. Conf. BEyond time and errors*, 2008, pp. 1–8.

[36] M. Shimabukuro, E. Flores, M. de Oliveira, and H. Levkowitz, "Coordinated views to assist exploration of spatio-temporal data: a case study," in *Proc. Int. Conf. Coordinated & Multiple Views in Exploratory Visualization*, 2004, pp. 107–117.

[37] C. Plaisant, "The challenge of information visualization evaluation," in *Proc. working conf. Advanced visual interfaces*, 2004, pp. 109–116.

[38] O. Hoeber and X. D. Yang, "Exploring web search results using coordinated views," in *Proc. Int. Conf. Coordinated & Multiple Views in Exploratory Visualization*, 2006, pp. 3–13.

[39] A. Perer and B. Shneiderman, "Balancing systematic and flexible exploration of social networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, pp. 693–700, 2006.

[40] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Phys Rev E*, vol. 69, p. 066133, 2004.

[41] D. Zajic, B. J. Dorr, J. Lin, and R. Schwartz, "Multi-candidate reduction: Sentence compression as a tool for document summarization tasks," *Inf. Process. Manage.*, vol. 43, pp. 1549–1570, 2007.

[42] D. R. Radev, P. Muthukrishnan, and V. Qazvinian, "The ACL anthology network corpus," in *Proc. ACL Workshop on Natural Language Process. and Inform. Retrieval for Digital Libraries*, Singapore, 2009.

[43] C. C. Kuhlthau, "Inside the search process: Information seeking from the user's perspective," *JASIS*, vol. 42, pp. 361–371, 1991.