# Serving the Readers of Scholarly Documents: A Grand Challenge for the Introspective Digital Library

(Invited Paper)

Min-Yen Kan
Web, Information retrieval / Natural language processing Group (WING)
School of Computing
National University of Singapore
http://www.comp.nus.edu.sg/~kanmy

***Abstract*—The scholarly literature produced by human civilization will soon be considered small data, able to be portably conveyed by the network and carried on personal machines. This semi-structured text centric knowledge base is a focus of attention for scholars, as the wealth of facts, facets and connections in scholarly documents are large. Such machine analysis can derive insights that can inform policy makers, academic and industrial management, as well as scholars as authors themselves.**

**There is another under-served community of scholarly document users that has been overlooked: the readers themselves. We call for the community to put more efforts towards supporting our own scholars (especially beginning scholars, new to the research process) with services using information retrieval and natural language processing technologies. Techniques that mine information from within the full text of a document could be used to introspect a digital library's materials, inferring better search metadata, improving scholarly document recommendation, and aiding the understanding of the text, figures, presentations and citations of our scholarly literature. Such an introspective digital library will enable scholars to assemble an understanding of other scholars' work more efficiently, and provide downstream machine reading applications with input for their analytics.**

The advent of big data has provided scholars the opportunity to instrument and measure the physical and social worlds on an unprecedented scale. As of 2014, per day, climatic and atmospheric science centers generate terabytes of data, while the users of Facebook and Twitter create social media content on the order of billions of messages. Creating, collecting, indexing and transmitting these amounts of data are clear scalability problems that need to be addressed. Big data and the services around data have brought about many new initiatives in computing, even at this present nascent period.

Perhaps even more urgent is in the sense-making of these raw primary data. The job title of "data scientist" has been half-jokingly referred to "data janitor", as the filtering, cleaning, re-formatting, and sense-making processes on these primary sources requires significant manual work [1]. Also, often times insights are found after linking disparate data sources together (i.e., people names in two different datasets), but such reference problems themselves are often difficult to solve, even with manual effort. Privacy concerns and legal rights are also significant issues that have yet to be adequately addressed and made actionable to safeguard the individual citizens in big data scenarios. These curatorial and techno-societal problems are mounting, and will likely need to be addressed before big data will reveal its promise.

Yet, much of what scholars discover are ultimately recorded in a secondary, digested form: the scholarly document. Relative to primary raw data, the scale of these secondary scholarly documents is small. In 2013, PubMed – a central source for medical literature funded by the U.S. government – registered about 1.1 million new abstracts [2], equivalent to about four thousand abstracts daily. While not close to the sum total of daily scientific knowledge created, it is not erroneous to cap current human knowledge production to tens of thousands of scholarly works per day, a far cry (as a conservative estimate, at least five magnitudes) from today's big data.

> While astronomers often have access to efficient and robust mechanisms that serve to archive, curate, and make primary data available. But very few parallel systems exist for derived data. Because most, if not all, scientific articles in astronomy are based on derived data, making such data visible, intelligible and available to the public is of fundamental importance — Pepe *et al.* [3].

While the big data of the primary empirical evidence is of import, it is also clear that the relatively small data of scholarly documents can be much better served. We would extend Pepe *et al.*'s reasoning to urge the community to study our own scholarly literature more deeply to enable better access. Compared with big data, scholarly documents are expert-reviewed and deeply self-curated: as a community, we invest authors', editors', reviewers' and typesetters' time to carefully vet the quality of our manuscripts and the metadata and linkages assigned to them. Also, the scale of scholarly documents is not trivial – small, yet manageable: we can imagine even today the machine processing of a few hundred documents per hours to be within the processing reach of individual graduate students and certainly faculty members' laboratories. These positive signals point to a clear research agenda of operationalizing the next generation digital library, what we term the *introspective* digital library (IDL).

In the following, we first describe the state of the current era of digital libraries, reviewing the current state-of-the-art of their machine processing. We then define the introspective digital library, a library that semantically understands the text of its collection and can facilitate its use for its readership. We propose several challenges that an introspective digital library should address, and conclude by showing that text processing, while a foundation technology for enabling the IDL, will need to incorporate with multimedia processing to understand its content, and eventually link back to the readers as an aggregate to prove its impact.

## I. Today: Electronic Digital Libraries

The digital library of today is the electronic library: where the focus is creating and ensuring proper access to the scholarly literature on demand. Digitization efforts and e-publishing have made the majority of the relevant (i.e., currently cited) scholarly documents available in electronic copy. The transition from the automated library[1] to the electronic one has been wildly successful: the pervasiveness of the anytime scholarly document has significantly raised expectations of scholars, especially those new to research. Scholarly work encumbered by access restrictions of either physical or digital nature, only find a limited audience.

Without loss of generality, a common workflow that many users follow is to use a search system to locate electronic scholarly documents of interest and consume them outside of the digital library (i.e., by printing them or reading a downloaded electronic copy). Search is largely keyword driven, using paper metadata — its title and in some cases, its abstract — as the source text to match a user's query against.

While the scholarly document is mostly text, with the exception of the title and abstract, the current digital library systems treat their documents as atomic objects, opaque, characterized by author provided metadata. From this vantage, the digital library is equivalent to a database of atomic records, where the objects in the collection itself cannot be scrutinized in further detail.

### A. Citations

The singular success where scholarly documents have been made transparent is the exception of bibliographic references. Eugene Garfield's effort in the 1950's led to his development of the longstanding citation indexing system, the *Science Citation Index* (SCI), and spurred other similar indices (*e.g.*, Scopus and Google Scholar). Exposing the bibliographic references made references actionable, creating a directed network of scholarly works. The resultant citation network allowed the aggregation of citations across documents, enabling measures of scientific impact for both individual works but also on aggregates for authors, departments, institutions, research fields and countries. While admittedly imprecise, this data source enabled a level of analysis of the literature that significantly influenced workflows of scholars, policy makers and research management in indirectly deciding faculty advancement and research funding decisions.

While SCI is curated and manually linked, even exposing the bibliographic references of the scholarly document in its raw form has spurred significant downstream research that aim to automate or enhance the scholarly document. The exposed citation metadata and network also gives rise to new functionality enabled by its aggregation.

*Automation.* Automated linkage of extracted references allow scalable document processing to build the citation graph algorithmically, most notably in Google Scholar[2], but also in open source platforms that support specific communities and hobbyists [4].

*Enhancement and Correction.* The references of a document represent just a fraction of related work. Due to space limitations, ignorance or overt omission, many relevant works are not cited by a source document. By inspecting properties of the actual citation graph, the link prediction task can recover and relate potentially relevant missing citations to a document. Inspecting the citation graph can also help to fix incorrect metadata, as incorrect conflating and splitting of names (especially for author names, but also names of publication venues, article and journal titles) can sometimes be corrected with holistic evidence from the citation graph.

Also, while citation counting is easily understood, there are many limitations of a raw count. Citations of a work from an impactful work versus an insignificant one arguably vary in their influence, but raw counts do not account for this factor. This can be posed as a recursive statement, resulting in the well-known PageRank algorithm [5], formerly a key component of webpage ranking in the early Google search engine. Running PageRank or its many variants has been shown to better assign influencer weights within such graphs (*e.g.*, [6]).

*Functionality.* Citation indices induced from aggregating the bibliographic references also have the useful property of being time-ordered. This can help scholars go back to find the source of an idea (by tracing backwards through references, called *backward chaining*) and find works that build upon a current work (by using a citation index to find newer works, or *forward chaining*).

Shared citations and references can also help enhance document search by providing an alternate means for computing document similarity. *Bibliographic coupling* [7] (where two documents share common references) and *co-citation analysis* [8] (where two documents are cited by a common work) can provide additional evidence of two documents' similarity through their relationship with a third party. When common citations and references are strong, a system can infer that two documents are strongly related even if their titles and terminology differ.

Enhancing document similarity computation serves to enhance many other downstream digital library functions that use similarity as a base. Similarity can be used with *documents* as targets for document search and paper recommendation and alerts, but also for *authors* (suggesting collaborators, reviewers or peer evaluators), for *venues* (suggesting potential venues for publications, given an input title or document text), and for *topics and keywords* (as found in scholarly works, useful in query expansion and expert finding). However useful, these applications all target external use of the contents of the digital

[1] Where only search and other services are automated, but physical copies are circulated.

[2] http://scholar.google.com.sg/intl/en/scholar/inclusion.html#indexing

library – i.e., they may help in searching, organizing or ranking the scholarly works, but do nothing to help use the knowledge within a work. For such capabilities, we need to introspect the contents of the scholarly work itself.

## II. Tomorrow: Introspective Digital Libraries

As scholarly works are now often authored digitally, there is little technological barrier to exposing the data and letting machine processes make sense of the data. With more scholarly data appearing in a freely available form on the web (about 30%, according to [9]), we are already in the transition towards the introspective digital library.

**Definition** *(Introspective Digital Library)* A digital library that is semantically able to understand the contents of its collection and act on its understanding to facilitate knowledge discovery, use and synthesis.

Introspective digital libraries (IDLs) differentiate from electronic ones in that they are able to act on their interpretation of the contents of their collection. There a few points worthy to detail more precisely. First, "understanding" connotes deep knowledge of the document, that goes beyond the simple inventory of word tokens used in the collection; it suggests that the IDL either infers or has access to knowledge (metadata) about the documents that is domain-specific. Second, that such knowledge not only facilitates discovery (as is done by search and indexing systems) but enables "use and synthesis": supporting the multiple reading phases of use (familiarization, understanding, interpretation and analysis), and the later synthesis stages of knowledge formation (comparing, re-implementing, evaluating, writing). We argue that such reader-centric services will be a critical component that will differentiate the introspective digital library from its electronic predecessor. Such functionality need not be limited *manual* use and synthesis, but also encompasses the notion of automation; i.e., machine processing could iteratively build on facts and assertions gleaned from its holdings to automatically discover new knowledge.

How do we get to the IDL? The first step is to make the scholarly documents accessible. Accessibility of the full document's text is mandatory prerequisite to reaching IDL functionality. Such accessibility improves inherently improves the current DL services. For example, document similarity via citation analysis can be supplemented with shared vocabulary analyses on full documents [10]. Full text analysis also makes for a better source for paper recommendation, especially when linking in scholarly document text from citing and referenced documents [11], [12].

To reach the IDL, we have to go further than just refining existing discovery and citation-based services. Full text, structure and figure accessibility will enable the critical functionality of an IDL. However, the prerequisite task of obtaining the document structure from the raw semi-structured text turns out to be a serious challenge. While publishers generate a logical, semantic representation of the published text — with notations such as figure and table captions, titles, affiliation, footnotes, etc. — which is usually saved in XML, the conventions are not consistently applied and vary even within a single publication venue due to adoption of stylistic conventions.

Inferring and making the logical document structure explicit from the formatted text has been shown to be a non-trivial process. It needs to be an automated process to be scalable. Information extraction workflows which infer such document structure [4], [13] and associate affiliations and authors [14] lay the necessary foundation to distinguish body text from captions, footnotes and page numbering. With the raw text differentiated by their logical document function, the introspective digital library can access the relevant text (*sans* other irrelevant raw text) to build semantic services for the document.

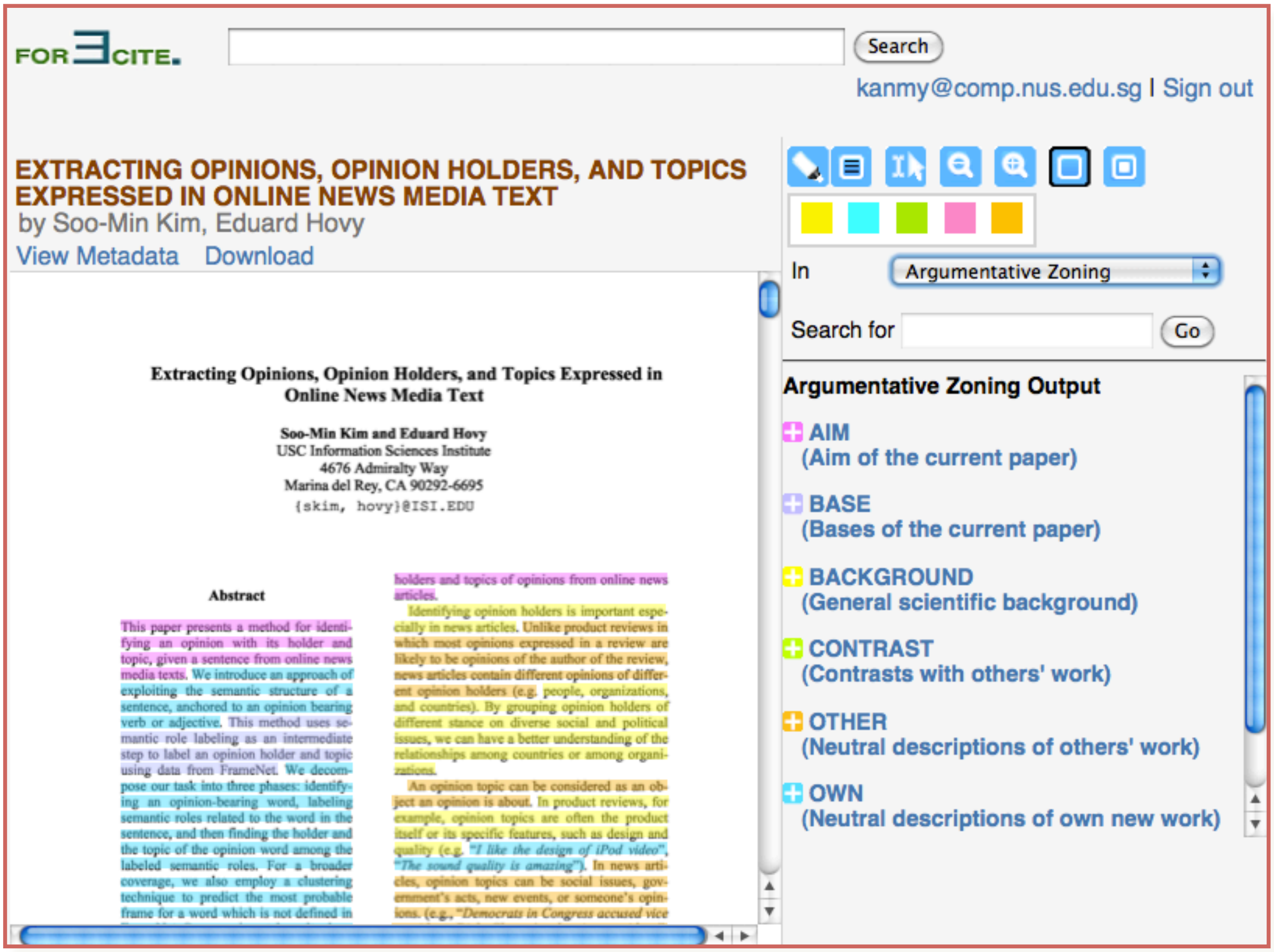
The running text of the scholar's argument in a work is the key text that will drive the IDL. As there are many ways that readers consume scholarly text, there are correspondingly many applications that can help to support the various use cases; we describe particular applications that we are working towards in our research, especially with the ForeCite prototype IDL [15]. For ease of exposition, we organize our discussion around the source data that they use from the full text: the running body text, citation sentences and figures.

### A. Body Text

*Keyphrases and Metadata.* A small step towards understanding a work is to distinguish its key terminology. Statistical topic models and keyphrase extraction (e.g., [16]) can help automatically extract key phrases from documents, to enrich its document representation. These can be used to supplement or replace author-provided keyword metadata, which typically lack coverage due to both time and space constraints.

Such models are often trained to be generic, so as to be widely applicable, but more fine-grained domain-specific aspects are useful to mine. For example, within the evidence-based medicine (EBM) domain, the PICO elements of Patient, Intervention, Comparison and Outcome are important aspects that characterize a work. Prospective readers of a EBM article – busy nurses and doctors – will want to ensure that the patient demographics in a study match the patients under their care, before committing time to read a study's findings. Machine inference of such metadata is valuable as it enables semantic search with both active and passive modes ("filter articles by patient age" and "articles are displayed sorted by patient cohort age") [17], [18].

*Symbolic Knowledge.* In certain disciplines (e.g., chemistry, mathematics), domain knowledge also has structure and is coded symbolically. Identifying, extracting, indexing and reasoning over symbolic representations also help build services [19]. In the domain of mathematics, mathematical equations are of significant import, but often only their natural language form is used for searching (i.e., people will search using "Pythagorean theorem" but not "$a^2 + b^2 = c^2$") [20]. IDLs that understand the duality of the knowledge representation can leverage this to perform their services better (e.g., also rank documents with the equivalent equation highly). Symbolic knowledge can also be used as a separate pathway to suggest relevance: with a math search engine that understands an equation's structure, we can cluster documents that use variant formulations of a central equation; i.e., contrast methods that have different formulations of iterative PageRank.



EXTRACTING OPINIONS, OPINION HOLDERS, AND TOPICS EXPRESSED IN ONLINE NEWS MEDIA TEXT
by Soo-Min Kim, Eduard Hovy

Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text

Soo-Min Kim and Eduard Hovy
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{skim, hovy}@ISI.EDU

Abstract

This paper presents a method for identifying an opinion with its holder and topic, given a sentence from online news media texts. We introduce an approach of exploiting the semantic structure of a sentence, anchored to an opinion bearing verb or adjective. This method uses semantic role labeling as an intermediate step to label an opinion holder and topic using data from FrameNet. We decompose our task into three phases: identifying an opinion-bearing word, labeling semantic roles related to the word in the sentence, and then finding the holder and the topic of the opinion word among the labeled semantic roles. For a broader coverage, we also employ a clustering technique to predict the most probable frame for a word which is not defined in

holders and topics of opinions from online news articles.

Identifying opinion holders is important especially in news articles. Unlike product reviews in which most opinions expressed in a review are likely to be opinions of the author of the review, news articles contain different opinions of different opinion holders (e.g. people, organizations, and countries). By grouping opinion holders of different stance on diverse social and political issues, we can have a better understanding of the relationships among countries or among organizations.

An opinion topic can be considered as an object an opinion is about. In product reviews, for example, opinion topics are often the product itself or its specific features, such as design and quality (e.g. *"I like the design of iPod video"*, *"The sound quality is amazing"*). In news articles, opinion topics can be social issues, government's acts, new events, or someone's opinions. (e.g., *"Democrats in Congress accused vice*

Argumentative Zoning Output

AIM
(Aim of the current paper)

BASE
(Bases of the current paper)

BACKGROUND
(General scientific background)

CONTRAST
(Contrasts with others' work)

OTHER
(Neutral descriptions of others' work)

OWN
(Neutral descriptions of own new work)

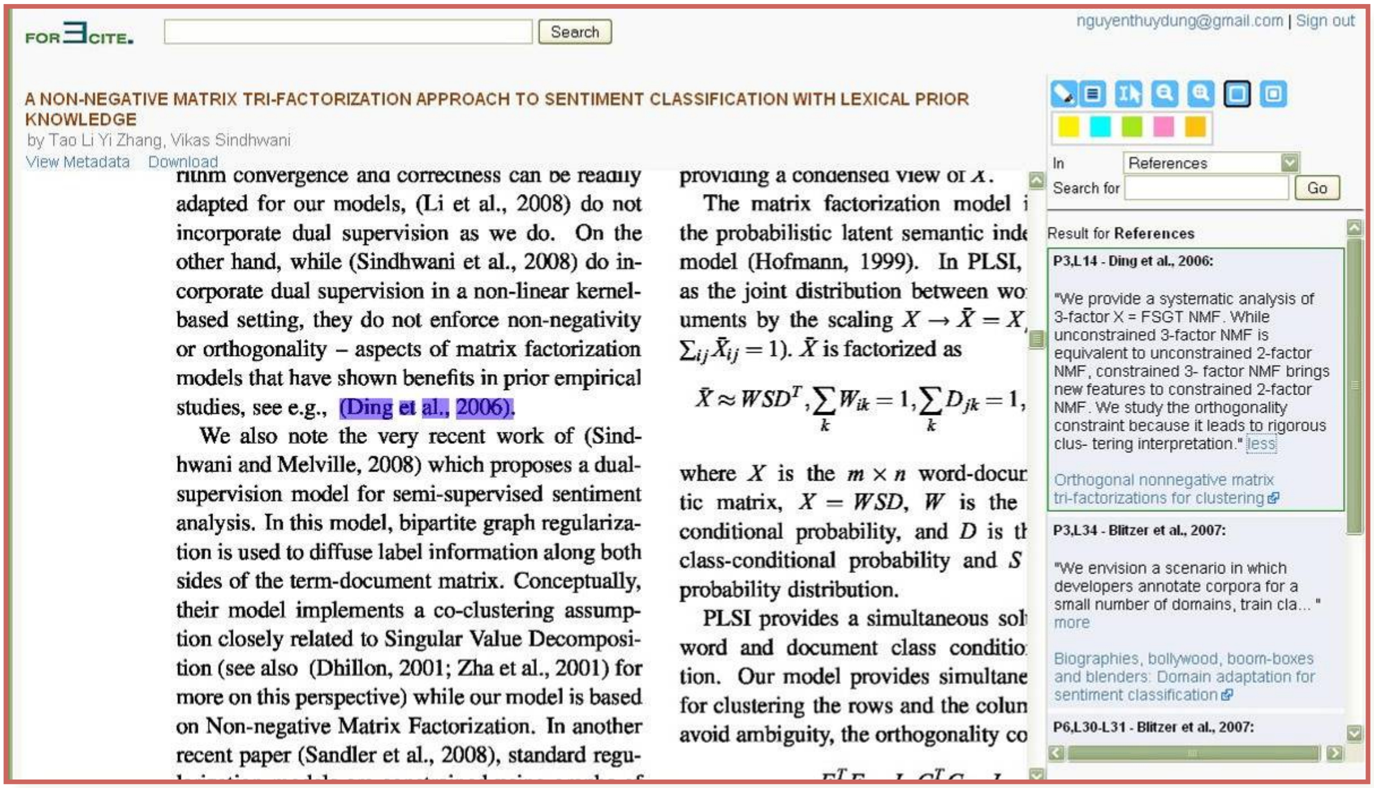
Fig. 1. Argumentative zoning sentence classification in the ForeCite prototype digital library. Sentences are color coded to indicate their rhetorical purpose within the document; the right panel allows a reader to browse through all sentences assigned to a particular argumentative zone.

*Argumentative Purpose.* While scholarly works often follow specific writing conventions, authors' own personal styles do vary substantially. Through the their phrasing, an authors marks the argumentative purpose of each sentence. Exposing such information in the IDL reading interface can can give readers an overview of the document's argumentative structure. Given an inventory of argumentative moves, this task can be structured as a supervised sentence classification task, known as argumentative zoning (AZ, [21]). Our work has shown that operationalizing this sentence classification system (as shown in Fig. 1) with reader-centric purposes (namely, *aim*, *background*, *basis*, *contrast*, *own*, *other* and *textual*) improves reader comprehension of key facts upon a first reading [22]. In particular, sentences identified as *Aim* sentences, help give a first-time reader a quick synopsis of the author attributed contributions of the work[3].

### B. *Citation Sentences*

While the identity of a work's bibliographic references are available in many DLs, they are only half of the full account of a citation. The remaining half is the text of citation itself; the text that attributes the reason for the reference. Such citation sentences (sometimes referred to as a "citance" [23]) can refine coarse-grained citation counts. With the evidence from the text of the accompanying citation sentences, an IDL can infer the intent of the citation, and possibly its semantic orientation. Citations exhibit different facets: in terms of scope (general, over the whole work, or specific to a paragraph, equation or section), sentiment (give positive endorsement, be neutral or show a weakness) [24], [25], as well as domain-specific purpose (e.g., in computer science, citing for a method, dataset, tool or evaluation metric). One might accord a reference to a work that is acknowledged specifically for enabling a target work more weight than one used as a prototypical example over a number of related works. Inferring such labels and then aggregating them provide a more accurate depiction of a work's impact.

[3] In contrast, the abstract often contains a condensed version of the entire paper's points, and may not describe all of the authors' goals.



A NON-NEGATIVE MATRIX TRI-FACTORIZATION APPROACH TO SENTIMENT CLASSIFICATION WITH LEXICAL PRIOR KNOWLEDGE
by Tao Li Yi Zhang, Vikas Sindhwani

rithm convergence and correctness can be readily adapted for our models, (Li et al., 2008) do not incorporate dual supervision as we do. On the other hand, while (Sindhwani et al., 2008) do incorporate dual supervision in a non-linear kernel-based setting, they do not enforce non-negativity or orthogonality – aspects of matrix factorization models that have shown benefits in prior empirical studies, see e.g., (Ding et al., 2006).

We also note the very recent work of (Sindhwani and Melville, 2008) which proposes a dual-supervision model for semi-supervised sentiment analysis. In this model, bipartite graph regularization is used to diffuse label information along both sides of the term-document matrix. Conceptually, their model implements a co-clustering assumption closely related to Singular Value Decomposition (see also (Dhillon, 2001; Zha et al., 2001) for more on this perspective) while our model is based on Non-negative Matrix Factorization. In another recent paper (Sandler et al., 2008), standard regu-

providing a condensed view of $X$.
The matrix factorization model i
the probabilistic latent semantic inde
model (Hofmann, 1999). In PLSI,
as the joint distribution between wo
uments by the scaling $X \to \bar{X} = X$,
$\sum_{ij} \bar{X}_{ij} = 1$). $\bar{X}$ is factorized as

$$\bar{X} \approx WSD^T, \sum_k W_{ik} = 1, \sum_k D_{jk} = 1,$$

where $X$ is the $m \times n$ word-docur
tic matrix, $X = WSD$, $W$ is the
conditional probability, and $D$ is tl
class-conditional probability and $S$
probability distribution.
PLSI provides a simultaneous sol
word and document class conditio
tion. Our model provides simultane
for clustering the rows and the colun
avoid ambiguity, the orthogonality co

P3,L14 - Ding et al., 2006:

"We provide a systematic analysis of 3-factor X = FSGT NMF. While unconstrained 3-factor NMF is equivalent to unconstrained 2-factor NMF, constrained 3- factor NMF brings new features to constrained 2-factor NMF. We study the orthogonality constraint because it leads to rigorous clus- tering interpretation." less

Orthogonal nonnegative matrix tri-factorizations for clustering

P3,L34 - Blitzer et al., 2007:

"We envision a scenario in which developers annotate corpora for a small number of domains, train cla... " more

Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification

P6,L30-L31 - Blitzer et al., 2007:

Fig. 2. Citation provenance at work. The claim indicated in the citation sentence on the left column in the reader panel is supported by the text on the right inset drawn from the cited document.

*Citation Provenance.* Tracing the provenance of a claim or idea is also possible with the understanding of citations. This refines the coarse-grained backward chaining strategy, with evidence from the citation sentence. For citations that can be inferred as being specifically scoped to a particular portion of a referenced work, the IDL can present the relevant section directly in the DL interface, allowing more seamless reading experience, as shown in Fig. 2.

*Document Summarization.* A citation sentence often describes the referenced work in relation to the target article, which we can think of as a focused summary. When a work has sufficient citations, aggregating such citation sentences produces a summary of the scholarly work from the perspective of the community. These are notably in complement with the work's abstract, which can be viewed is a summary from the authors' perspective. Taken together with body text understanding, automatic summarization can leverage these many sources to create summaries of varying length, suitable for use in different contexts.

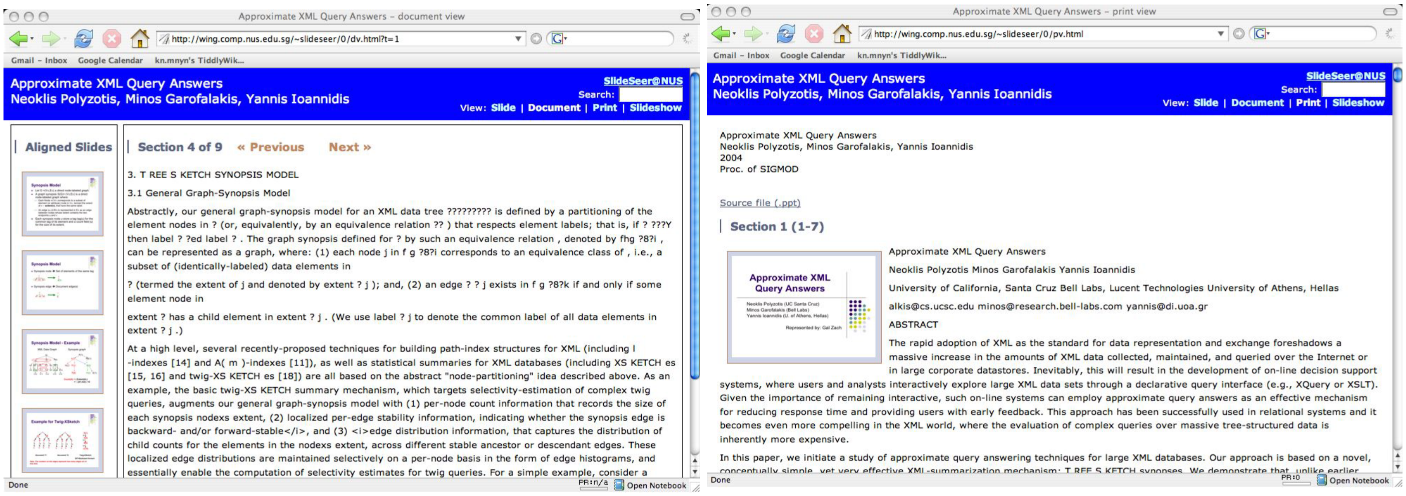
Summarization of multiple-related documents is arguably more fruitful than for single documents. Automatic survey paper generation [26] and related work summarization [27] make use of information derived from multiple documents to both generate and extract text to form comparative summaries. As the component systems of domain-specific metadata extraction, argumentative zoning and citation understanding improve, more useful and fine-grained automatic summaries of scholarly documents will begin to have impact on scientific processes.

### C. Figures

Scholarly documents are increasingly composed of multimedia. In many fields, figures and tables serve as windows into the primary data of a work, as aggregates or samples. An IDL will need to extend its services to be sensitive to all of its different forms of embedded media.

*Cross Reference Detection and Scoping.* While figures are stylistically often interpretable with only the help of a caption, they are also usually cross-referenced explicitly and discussed in the document's body text. However, inferring when the discussion veers from the figure's topic to other topics has yet to be well explored. Techniques that utilize prior research on citation [28] or co-reference scoping may provide good baselines. Also, some mentions do not follow explicit conventions, so automated detection can help build the necessary linkages when only implicit (e.g., "As shown on the previous page, the subfigure shows ..."). Inferring automatic linkages will enrich the representation of such multimedia objects via their accompanying text, which can give sizable performance over purely content-based solutions (*i.e.*, Web image search relies heavily on nearby text to characterize it) which ignore the context.

*Data Extraction from Figures.* Table and chart understanding can help recover primary data points from charts. The document understanding community has been developing these capabilities independently, establishing methods to parse information from axes and legends to recover quantitative amounts (e.g., [30], [31]) from scientific figures. Given adequate cross



Approximate XML Query Answers
Neoklis Polyzotis, Minos Garofalakis, Yannis Ioannidis

Section 4 of 9 « Previous Next »

3. T REE S KETCH SYNOPSIS MODEL

3.1 General Graph-Synopsis Model

Abstractly, our general graph-synopsis model for an XML data tree ????????? is defined by a partitioning of the element nodes in ? (or, equivalently, by an equivalence relation ?? ) that respects element labels; that is, if ? ???Y then label ? ?ed label ? . The graph synopsis defined for ? by such an equivalence relation , denoted by fhg ?8?i , can be represented as a graph, where: (1) each node j in f g ?8?i corresponds to an equivalence class of , i.e., a subset of (identically-labeled) data elements in

? (termed the extent of j and denoted by extent ? j ); and, (2) an edge ? ? j exists in f g ?8?k if and only if some element node in

extent ? has a child element in extent ? j . (We use label ? j to denote the common label of all data elements in extent ? j .)

At a high level, several recently-proposed techniques for building path-index structures for XML (including I -indexes [14] and A( m )-indexes [11]), as well as statistical summaries for XML databases (including XS KETCH es [15, 16] and twig-XS KETCH es [18]) are all based on the abstract "node-partitioning" idea described above. As an example, the basic twig-XS KETCH summary mechanism, which targets selectivity-estimation of complex twig queries, augments our general graph-synopsis model with (1) per-node count information that records the size of each synopsis nodexs extent, (2) localized per-edge stability information, indicating whether the synopsis edge is backward- and/or forward-stable</i>, and (3) <i>edge distribution information, that captures the distribution of child counts for the elements in the nodexs extent, across different stable ancestor or descendant edges. These localized edge distributions are maintained selectively on a per-node basis in the form of edge histograms, and essentially enable the computation of selectivity estimates for twig queries. For a simple example, consider a

Approximate XML Query Answers – print view
http://wing.comp.nus.edu.sg/~slideseer/0/pv.html
Gmail – Inbox
Google Calendar
kn.mnyn's TiddlyWik...
SlideSeer@NUS
Search:
View: Slide | Document | Print | Slideshow
Approximate XML Query Answers
Neoklis Polyzotis (UC Santa Cruz)
Minos Garofalakis (Bell Labs)
Yannis Ioannidis (U. of Athens, Hellas)
Represented by: Gal Zach
Done
PR:0
Open Notebook

Approximate XML Query Answers
Neoklis Polyzotis, Minos Garofalakis, Yannis Ioannidis

Approximate XML Query Answers
Neoklis Polyzotis, Minos Garofalakis, Yannis Ioannidis
2004
Proc. of SIGMOD

Source file (.ppt)

Section 1 (1-7)

Approximate XML Query Answers

Neoklis Polyzotis Minos Garofalakis Yannis Ioannidis

University of California, Santa Cruz Bell Labs, Lucent Technologies University of Athens, Hellas

alkis@cs.ucsc.edu minos@research.bell-labs.com yannis@di.uoa.gr

ABSTRACT

The rapid adoption of XML as the standard for data representation and exchange foreshadows a massive increase in the amounts of XML data collected, maintained, and queried over the Internet or in large corporate datastores. Inevitably, this will result in the development of on-line decision support systems, where users and analysts interactively explore large XML data sets through a declarative query interface (e.g., XQuery or XSLT). Given the importance of remaining interactive, such on-line systems can employ approximate query answers as an effective mechanism for reducing response time and providing users with early feedback. This approach has been successfully used in relational systems and it becomes even more compelling in the XML world, where the evaluation of complex queries over massive tree-structured data is inherently more expensive.

In this paper, we initiate a study of approximate query answering techniques for large XML databases. Our approach is based on a novel, conceptually simple, yet very effective XML-summarization mechanism: T REE S KETCH synopses. We demonstrate that, unlike earlier

Fig. 3. SlideSeer's [29] coordinated document view and print views. In the document view, the accompanying slide presentations' context is given on the left, along with the paper's text (re-flowed).

reference scoping functionality, the accompanying body text may serve to constrain the extraction process and enhance performance. Extracted data can then be used to comparisons across multiple works, which can be another source of input for multi-document summarization.

*Presentation Alignment.* In certain fields, scholarly works are also presented orally and visually, independent of the document artifact. Visual artifacts, such as slide presentations, offer a dual representation of a work that can serve as a summary or enhancement of a work. Locating, indexing and linking such auxiliary artifacts to the primary scholarly work assist in rendering a complete picture for a reader. As with the main tenet of the IDL, such artifacts should be aligned to the work at a fine-grained resolution, allowing deep linkage between the presentation and work. The area of multimedia research has pioneered this field (e.g., [32]) as well as providing commercial products[4]. The SlideSeer prototype [29], also shown in Fig. 3 casted this as the alignment task between two text streams, while later work [33] showed the importance of visual features in tackling this problem, especially on visually-salient slides which contain little text (e.g., those that feature figures). With a corpus of aligned presentations and documents, the task of presentation generation [34] becomes a target, which can be thought of generating a single document multimodal summary with specific constraints.

With these directions, we see that introspection clearly extends to multimedia. Similar to symbolic knowledge, an understanding of these multimedia artifacts and their meaning will enable better content-based representations of the scholarly work.

## III. The Challenges Ahead

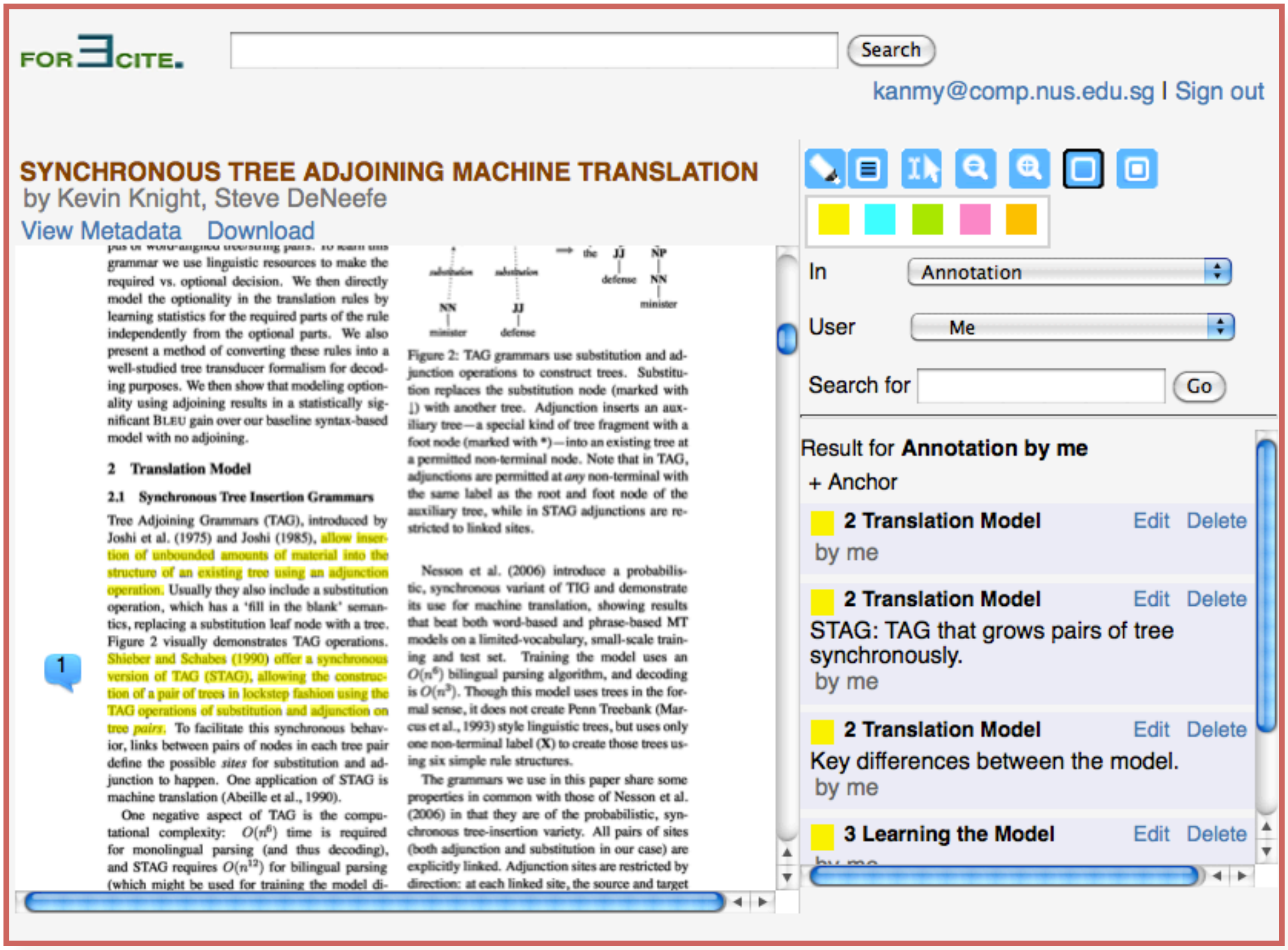
We have focused primarily on tasks that assist the reader with the current document at hand. We are particularly interested with services that help beginning scholars consume and sense-make from the such documents. Implementing interfaces that affords distraction-free reading while maintaining easy access to auxiliary functions (e.g., notetaking, highlighting, context switching between auxiliary and linked work) is a trade-off that will need to be managed well. Information needs vary widely across disciplines and even within a discipline, the use of the scholarly document varies with the lifecycle stage of research a scholar is in (e.g., ideation, problem formulation, discovery, reading for breadth, reading for depth, implementation, comparison, writing, and presenting). With a deeper understanding of its content, IDLs are poised to invent new services to augment their scholars' investment in reading. The grand challenge of the introspective digital library is to imagine and implement these services that will facilitate more effective scholarly communication.

We also note that the digital library of today lives in the era of user generated content (UGC; also known as Web 2.0). This environment should influence DL architecture by placing the reader (rather than fellow authors) as an important authoring stakeholder. Aggregating the activities of the many readers of a scholarly document will enable "collaborative filtering" like services. For example, aggregating the textual highlights of many readers of a single document can provide a "heat map" visualization of the important portions of a document (similar to Fig. 4). We believe UGC's value proposition to the IDL is in its lightweight contribution framework: User defined tags and simple conventions (such as #hashtagging) will allow users to create the functionality they want in their view of their scholarly interests, while having the ulterior purpose of training autonomous DL services to contribute such tags automatically.

A final aspect of the challenge will be to integrate the services mentioned here with larger family of services that other developments in the digital library space (both academic and commercial) are pioneering. Building the appropriate infrastructure to federate the documents as well as the services will be a key obstacle before the widespread implementation of the introspective digital library is to occur [35].

In closing, we note that many of today's netizenry consume news natively on news portals – not only reading, but in cases, also actively contributing to the discussion. However, the same cannot be said for scholarly work – most scholars

[4] e.g., StreamSage's synchronization software.



SYNCHRONOUS TREE ADJOINING MACHINE TRANSLATION
by Kevin Knight, Steve DeNeefe

grammar we use linguistic resources to make the required vs. optional decision. We then directly model the optionality in the translation rules by learning statistics for the required parts of the rule independently from the optional parts. We also present a method of converting these rules into a well-studied tree transducer formalism for decoding purposes. We then show that modeling optionality using adjoining results in a statistically significant BLEU gain over our baseline syntax-based model with no adjoining.

2 Translation Model

2.1 Synchronous Tree Insertion Grammars

Tree Adjoining Grammars (TAG), introduced by Joshi et al. (1975) and Joshi (1985), allow insertion of unbounded amounts of material into the structure of an existing tree using an adjunction operation. Usually they also include a substitution operation, which has a 'fill in the blank' semantics, replacing a substitution leaf node with a tree. Figure 2 visually demonstrates TAG operations. Shieber and Schabes (1990) offer a synchronous version of TAG (STAG), allowing the construction of a pair of trees in lockstep fashion using the TAG operations of substitution and adjunction on tree *pairs*. To facilitate this synchronous behavior, links between pairs of nodes in each tree pair define the possible *sites* for substitution and adjunction to happen. One application of STAG is machine translation (Abeille et al., 1990).

One negative aspect of TAG is the computational complexity: $O(n^6)$ time is required for monolingual parsing (and thus decoding), and STAG requires $O(n^{12})$ for bilingual parsing (which might be used for training the model di-

Figure 2: TAG grammars use substitution and adjunction operations to construct trees. Substitution replaces the substitution node (marked with ↓) with another tree. Adjunction inserts an auxiliary tree—a special kind of tree fragment with a foot node (marked with *)—into an existing tree at a permitted non-terminal node. Note that in TAG, adjunctions are permitted at *any* non-terminal with the same label as the root and foot node of the auxiliary tree, while in STAG adjunctions are restricted to linked sites.

Nesson et al. (2006) introduce a probabilistic, synchronous variant of TIG and demonstrate its use for machine translation, showing results that beat both word-based and phrase-based MT models on a limited-vocabulary, small-scale training and test set. Training the model uses an $O(n^6)$ bilingual parsing algorithm, and decoding is $O(n^3)$. Though this model uses trees in the formal sense, it does not create Penn Treebank (Marcus et al., 1993) style linguistic trees, but uses only one non-terminal label (**X**) to create those trees using six simple rule structures.

The grammars we use in this paper share some properties in common with those of Nesson et al. (2006) in that they are of the probabilistic, synchronous tree-insertion variety. All pairs of sites (both adjunction and substitution in our case) are explicitly linked. Adjunction sites are restricted by direction: at each linked site, the source and target

Result for **Annotation by me**

+ Anchor

🟨 **2 Translation Model** Edit Delete
by me

🟨 **2 Translation Model** Edit Delete
STAG: TAG that grows pairs of tree synchronously.
by me

🟨 **2 Translation Model** Edit Delete
Key differences between the model.
by me

🟨 **3 Learning the Model** Edit Delete

Fig. 4. Collaborative highlights and annotations in the ForeCite prototype digital library. Anonymous readers textual highlights are aggregated and available to all readers, while annotated comments can be public or restricted to be shared with specific user groups.

still consume such knowledge outside of a digital library framework. A sign that we have completed the grand challenge of the introspective digital library, is when the IDL becomes an essential element of a scholar's toolset; when we can no longer imagine scholarship without the augmentation afforded by the IDL.

## Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

The author would also like to acknowledge the many members of the Web, Information Retrieval / Natural Language Processing Group (WING) at the National University of Singapore that have been involved in the ForeCite prototype digital library research, including Bamdad Bahrani, Muthu Kumar Chandrasekaran, Philip S. Cho, Dinh-Trung Dang, Huy Hoang Do, Markus Hänse, Ching Hoi Andy Hong, Minh-Thang Luong, Jesse Prabawa Gozali, Thuy Dung Nguyen, Kazunari Sugiyama and Yee Fan Tan.

## References

[1] S. Lohr, "For big-data scientists, janitor work is key hurdle to insights," http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html, August 2014.

[2] A. D. Corlan, "Medline trend: automated yearly statistics of pubmed results for any query," http://dan.corlan.net/medline-trend.html, 2004.

[3] A. Pepe, A. Goodman, A. Muench, M. Crosas, and C. Erdmann, "How do astronomers share data? reliability and persistence of datasets linked in aas publications and a qualitative study of data practices among us astronomers," *PLoS ONE*, vol. 9, no. 8, August 2014.

[4] I. G. Councill, C. L. Giles, and M.-Y. Kan, "ParsCit: an open-source crf reference string parsing package." in *Proceedings of the Language Resources and Evaluation Conference (LREC '08)*, Marrakesh, Morrocco, May 2008.

[5] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, pp. 107–117, 1998.

[6] J. Bollen, M. A. Rodriguez, and H. V. de Sompel, "Journal status," *CoRR*, vol. abs/cs/0601030, 2006. [Online]. Available: http://arxiv.org/abs/cs/0601030

[7] M. M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation*, vol. 14, no. 1, pp. 10–25, 1963.

[8] H. Small, "Co-citation in the scientific literature: a new measure of the

relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, pp. 265–269, 1973.

[9] Z. Wu, J. Wu, M. Khabsa, K. Williams, H.-H. Chen, W. Huang, S. Tuarob, S. R. Choudhury, A. Ororbia, P. Mitra *et al.*, "Towards building a scholarly big data platform: Challenges, lessons and opportunities," in *Digital Libraries 2024*. ACM, 2014.

[10] K. Williams, J. Wu, and C. L. Giles, "Simseerx: A similar document search engine," in *Proceedings of the 2014 ACM Symposium on Document Engineering*, ser. DocEng '14. New York, NY, USA: ACM, 2014, pp. 143–146. [Online]. Available: http://doi.acm.org/10.1145/2644866.2644895

[11] K. Sugiyama and M.-Y. Kan, "Scholarly paper recommendation via user's recent research interests," in *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 2010, pp. 29–38.

[12] ——, "A comprehensive evaluation of scholarly paper recommendation using potential citation papers," *International Journal on Digital Libraries*, pp. 1–19, 2014.

[13] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp, "Evaluation of header metadata extraction approaches and tools for scientific pdf documents," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2013, pp. 385–386.

[14] M.-T. Luong, T. D. Nguyen, and M.-Y. Kan, "Logical structure recovery in scholarly articles with rich document features," *Journal of Digital Library Systems. Forthcoming*, 2011.

[15] T. D. Nguyen, M.-Y. Kan, D.-T. Dang, M. Hänse, C. H. A. Hong, M.-T. Luong, J. P. G. K. Sugiyama, and Y. F. Tan, "Forecite: Towards a reader-centric scholarly digital library," in *JCDL*, 2010.

[16] M.-Y. K. Su Nam Kim, Olena Medelyan and T. Baldwin, "Automatic keyphrase extraction from scientific articles," *Language Resources and Evaluation*, December 2012.

[17] S. Lin, J.-P. Ng, S. Pradhan, J. Shah, R. Pietrobon, and M.-Y. Kan, "Extracting formulaic and free text clinical research articles metadata using conditional random fields," in *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*. Association for Computational Linguistics, 2010, pp. 90–95.

[18] C.-T. Tsai, G. Kundu, and D. Roth, "Concept-based analysis of scientific literature," in *Proceedings of the 22nd ACM international conference on Conference on information &#38; knowledge management*, ser. CIKM '13. New York, NY, USA: ACM, 2013, pp. 1733–1738. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505613

[19] P. Mitra, C. L. Giles, B. Sun, and Y. Liu, "Chemxseer: a digital library and data repository for chemical kinetics," in *Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience*. ACM, 2007, pp. 7–10.

[20] J. Zhao, M.-Y. Kan, and Y. L. Theng, "Math information retrieval: user requirements and prototype implementation," in *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2008, pp. 187–196.

[21] S. Teufel and M. Moens, "Summarizing scientific articles: experiments with relevance and rhetorical status," *Computational linguistics*, vol. 28, no. 4, pp. 409–445, 2002.

[22] S. Teufel and M.-Y. Kan, "Robust argumentative zoning for sensemaking in scholarly documents," *Advanced Language Technologies for Digital Libraries*, pp. 154–170, 2011.

[23] P. I. Nakov, A. S. Schwartz, and M. Hearst, "Citances: Citation sentences for semantic analysis of bioscience text," in *Proceedings of the SIGIR04 workshop on Search and Discovery in Bioinformatics*, 2004, pp. 81–88.

[24] D. Radev and A. Abu-Jbara, "Rediscovering ACL discoveries through the lens of ACL anthology network citing sentences," in *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Association for Computational Linguistics, 2012, pp. 1–12.

[25] A. Abu-Jbara, J. Ezra, and D. Radev, "Purpose and polarity of citation: Towards nlp-based bibliometrics," *Proceedings of NAACL-HLT*, pp. 596–606, 2013.

[26] R. Jha, R. Coke, and D. Radev, "Surveyor: A system for generating coherent survey articles for scientific topics," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI '15)*. AAAI, 2015.

[27] C. D. V. Hoang and M.-Y. Kan, "Towards automated related work summarization," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010, pp. 427–435.

[28] A. Abu-Jbara and D. Radev, "Reference scope identification in citing sentences," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 80–90.

[29] M.-Y. Kan, "SlideSeer: A digital library of aligned document and presentation pairs," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2007, pp. 81–90.

[30] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 235–242.

[31] W. Huang, C. L. Tan, and W. K. Leow, "Associating text and graphics for scientific chart understanding," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005, pp. 580–584.

[32] *2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Edinburgh, UK, 2005.

[33] B. Bahrani and M.-Y. Kan, "Multimodal alignment of scholarly documents and their presentations," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2013, pp. 281–284.

[34] Y. Hu and X. Wan, "PPSGen: Learning-based presentation slides generation for academic papers," in *Proceedings of the 21st international jont conference on Artifical intelligence (IJCAI '13)*. IEEE, 2013, pp. 2099–2105.

[35] K. Williams, J. Wu, S. R. Choudhury, M. Khabsa, and C. L. Giles, "Scholarly big data information extraction and integration in the CiteSeer$\chi$ digital library," in *2014 IEEE 30th International Conference on Data Engineering Workshops (ICDEW '14)*. IEEE, 2014, pp. 68–73.