Content-based citation analysis: The next generation of citation analysis

Ying Ding[1], Min Song[2], Xiaolong Wang[3], Guo Zhang[1], Chengxiang Zhai[3], Tamy Chambers[1]

[1]Department of Information and Library Science, School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA
[2]Department of Library and Information Science, Yonsei University, Seoul, South Korea
[3]Department of Computer Science, College of Engineering, University of Illinois, Urbana, Illinois, USA

Abstract

Traditional citation analysis has been widely applied to detect patterns of scientific collaboration, map the landscapes of scholarly disciplines, assess the impact of research outputs, and observe knowledge transfer across domains. It is, however limited, as it assumes all citations are of similar value and weights each equally. Content-based citation analysis (CCA) addresses a citation's value by interpreting each based on their contexts at both syntactic and semantic level. This paper provides a comprehensive overview of CAA research in terms of its theoretical foundations, methodical approaches, and example applications. In addition, we highlight how increased computational capabilities and publicly available full-text resources have opened this area of research to vast possibilities, which enable deeper of citation analysis, more accurate citation prediction, and increased knowledge discovery.

1. Introduction

The analysis of scholarly communication through citation patterns has been extensively used to detect scientific collaboration, map the landscapes of scholarly disciplines, assess the impact of research outputs, and observe knowledge transfer across domains. Papers and their citations have been used to form networks (e.g., paper citation networks, author citation networks, co-author networks, author co-citation networks, or journal co-citation networks) where nodes represent papers, authors, or journals, and edges indicate the number of times each has been cited, co-authored, or co-cited. While the simple counting of citations remains one of the most measurable indicators of research impact, it is limited as considers neither the location (e.g., where the reference has been mentioned in a citing article) nor the semantics (e.g., why the reference has been cited in a citing article) of a reference (see Figure 1).

The development of the Science Citation Index in the early 1960s ushered in the practice of citation analysis studies, which focuses on whom researchers cite, which documents they cite, and which journals they cite (Hjørland and Nielsen, 2001, Nicolaisen, 2007). Content-based citation analysis (CCA) is the next generation of citation analysis. It aims to expand upon the analysis of citation frequencies by using reference information at both syntactic (e.g., the position of where or which style a reference is mentioned in a citing article) and semantic level (e.g., how a reference is cited and how a knowledge concept or a domain entity is cited). In the late 60s and 70s, manual effort was required to address the "how and why" questions of citation analysis and, as a result, only small sets of papers were used in the studies. Although this allowed for the categorization and systematic analysis of citation motivations, the results were often not generalizable due to the limited sample size. However, recent developments in computing and

information services such as, full-text papers open to the public, computers capable of handling large-scale textual data, and innovative algorithms efficient at detecting both the location and semantic context surrounding a reference (Teufel, 2000), now offer the ability to ask the previous "how and why" questions of citation analysis on a larger scale. This has, in turn, called for a revisiting of the subject and the review presented here.

Figure 1: An example of citation content with location and semantic information

The primary goal of this paper is to provide an overview of CCA in terms of its foundations, approaches, and applications (see Figure 2). This paper is organized as follows: Section 2 provides the theoretical foundations for CCA; Section 3 describes the methodological approaches applied in CCA; Section 4 examines applications of CCA; and Section 5 provides a summary and identifies avenues for future research.

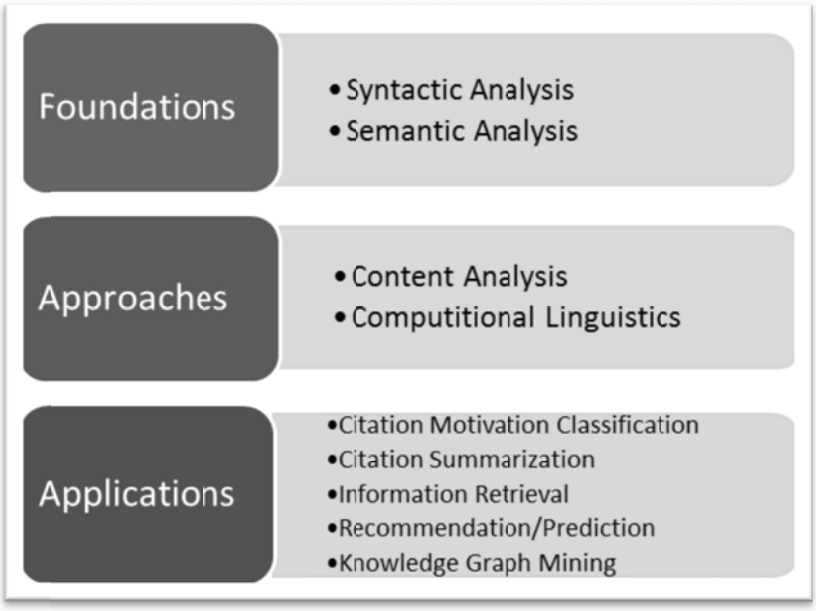

Figure 2: Content-based Citation Analysis (CAA) Overview

2. Theoretical Foundations

Scientific papers are the predominant means by which researchers promote their research findings. They are also the predominant means by which researchers garner attention (Cronin, 2005) and thereby acquire, through citation, peer recognition of their knowledge. Whether as a levied tax (Becher, 1989), or as footprints marking the passing of knowledge (Cronin, 1981), the citation has long been a subject of scholarly analysis. In this context, the main utility of CCA is the analysis of a citation's context within the full text of the scientific paper, rather than its simple frequency. As such, CCA can be divided into two tracks: syntactic CCA and semantic CCA.

2.1 Syntactic CCA

Over the long history of knowledge accumulation, scientific communities have established consistent formats for reporting knowledge in scientific papers. The current organization of information into the standard sections of introduction, related works, methods, results, discussion, and conclusions has been widely adopted by major journals and conferences. Based on a citation's location within a standardized section, it becomes possible to analyze a citation's perceived level of usefulness.

Voos and Dagaev (1976) first addressed the issue of treating all citations equally. They analyzed citation contribution based on its location within the introduction, methodology, discussion, or conclusion sections of a citing article. Finding, among other things, that the introduction contained more highly cited articles than other sections, they concluded that the contribution of a citation should be based on both its frequency and its location within the citing article. Herlach (1978) extended this argument by contending that a paper, cited in the introduction or literature review section and later again in the methodology or discussion section, should be regarded as having a greater contribution to the citing article than others, which may have been referenced only once in the entire citing article.

Peritz (1983) also used frequency to calculate a citation's contribution to the citing article; he, however, differentiated between formal citations, which reference the author name with the publication year and informal citations which indicate only the author's name. Bonzi (1982) similarly categorized citation relevance based on how citations were mentioned in the citing article. After analyzing 31 library and information science articles with nearly 500 references, she identified the following four citation styles: 1) those citations not specifically mentioned in the text (e.g., "several studies have dealt with …"), 2) those barely mentioned in the text (e.g., "Smith has studied the impact of"), 3) those with only one quotation or discussion of point (e.g., "Smith found that…"), and 4) those with two or more quotations or discussion points. Moravcsik and Murugesan (1975) studied the redundant patterns of citations in which an author cites several works simultaneously within a single citation block to indicate a list of examples, such as similar situations, good sources, or applications of classic methods. Their analyses of 575 references in 30 articles of theoretical high-energy physics found that one-third of references were redundant, while one-seventh were negational, and two-fifths were perfunctory. A later study showed that these textual and non-textual characteristics (e.g., number of references, figures, number of uncommon words, and the readability of abstracts) can account for 15% to 35% of the variation in citation frequencies (Snizek, Oehler, & Mullins, 1991).

It is important to note that these early studies of syntactic CCA were conducted manually on small paper sets. Later, however, Maričić, et al. (1998) conducted a citation analysis based on the location of references in more than 350 papers. Their results showed that the methods, results, and discussion sections contain more meaningful citations than the introduction section. Supporting this finding, Suppe (1998) explained that article sections about methods, data, and interpretations were important to the evaluation of whether the new findings could be integrated into the common knowledge base of a discipline.

2.2 Semantic CCA

The semantic relation between documents connected by citations has been discussed extensively for decades. As early as 1957, Merton (1957) claimed that citation was driven by the norms of science which he observed as a part of the compensating system for science. In his commentaries on the citation motivation of the authors, Garfield (1964) presented 15 reasons why authors cite other texts. Although his work relied on observation and anecdotes, which did not shed light on citation frequency, it was one of the first proposals to study citation motivation in depth. Near the same time, Lipetz (1965), while developing relational indicators to index documents, identified 29 categories describing relationships between cited articles and citing articles. She grouped these categories into the following four clusters: 1) original scientific contribution of the citing paper, 2) other than original scientific contribution of the citing paper, 3) relationship identification between the citing paper and the cited paper, and 4) scientific contribution of the cited paper to the citing paper. She proposed authors use these categories to clearly state their citing motivation for each citation in their article. While neither Garfield (1964) nor Lipetz (1965) provided empirical evidence of citing behavior, their commentaries, none-the-less, have been frequently used by researchers studying semantic citation characteristics.

Researchers in the late 1970s devoted much attention to author motivation in the examination of citation practice. Their goal was to understand if citation frequency could quantitatively measure author quality and prestige. Gilbert (1977) was the first to argue that citation served as the tool of persuasion, rather than evaluation, as was the "normative view" (p.113). Sociologists and bibliometricans have explored the finer structure of citation practice through different dimensions: examination of text surrounding citations (Chubin & Moitra, 1975), analysis of the function and quality of citations (Moravcsik & Murugesan,

1975), and identification of previous research use (Spiegel-Rösing, 1977). These researchers explored citation contribution extensively through labor-intensive and small-sized content analysis. Chubin and Moitra (1975) set up a tree hierarchy solely focusing on the different contribution levels of cited works: confirmative (four types) or negative (two types). Moravcsik and Murugesan (1975) proposed relationship indicators to distinguish between the critical and non-critical contributions of citations, such as whether the citing paper was extending previous ideas or proposing a new viewpoint, or denying or confirming the cited work. Spiegel-Rösing (1977) separated the methodological function of citation (e.g., providing data, developing methods, etc.) from the general function of citation (e.g., historical background of a subject domain) to conduct the first citation context analysis outside the field of physics. She found that 80% of citations substantiate a statement or point to further information, 5.8% were mentioned only in the introduction or literature review as the state-of-the-art, and 5.3% were used to compare data.

The first large-scale citation content analysis (Oppenheim & Renn, 1978) used 23 highly-cited old papers in the fields of physics and physical chemistry and found nearly 40% of all citations referenced the historical background. Small (1978) was among the first to study the scientific content of a citation by viewing it as a symbol of a concept or method, similar to Garfield's (1974) use of cited documents as subject headings in an indexing system. Small proposed that referencing was a process of labeling, and as such, citation context (i.e., the surrounding text of a citation) constituted the author's interpretation of the cited work. He examined the citation contexts (i.e., 2-3 sentences around the points where citations appeared) of citations within a set of highly cited articles in chemistry and found that most were not "research front papers" (p.334) but rather "well-established instructions on how to carry out certain basic operations at the lab bench or at the desk" (p.334). He thus concluded that highly cited articles act as symbolic exemplars.

In the 1980s while Small (1982) and Cronin (1984) both extensively studied the comparison of citation classification schemas; other researchers conducted postal surveys or in-person interviews on the topic of citing behavior. Representative works of the latter include the following: a semantic citation analysis to categorize the citation behaviors based on survey and interview (Hodges, 1972), multiple surveys of recently published authors in the field of chemistry (Brooks, 1985; Vinkler, 1987), a survey which employed the Moravcsik and Murugesan models (Cano, 1989), and two extensive-scale surveys of psychologists (Shadish, Tolliver, Gray, & Gupta, 1995). Conversely, McCain and Turner (1989), contending citation choice reflected the perceived usefulness of the cited work, conducted a manual bibliometric analysis of citation patterns within the field of Molecular Genetics. Focusing on the aging patterns of individual journal articles, they explored relationships between several content-related citation variables in eleven articles.

These early efforts sought to justify the feasibility of using citations to evaluate scholarly impact by classifying citation motivation and identifying the function of citations using a relatively limited set of articles (10-100 full-text articles). Methods were restricted to interview or manual analysis. Later, however, computer technology developments led to automatic data processing algorithms capable of massive content-based data analysis. Teufel, Siddharthan, and Tidhar (2006a, 2006b) proposed a reliable citation function annotation schema which, allowed a supervised machine learning algorithm to automatically classify citation functions (e.g., reasons that a researcher cites a particular paper) using both shallow and deep natural language features. In that schema, they used four top-level categories (explicit statement of weakness, contrast or comparison to another work, agreement/usage/compatibility to another work, and a neutral category) to label each citation. Avoiding sociologically oriented distinctions (e.g., paying homage to pioneers), they instead aimed for reliable annotation. A test of their approach on

360 conference articles found a strong relationship between citation function and sentiment classification. Small (2011) later analyzed citation sentiments using the text surrounding references in scientific papers and by combining science mapping with a linguistic analysis of the citation contexts to deepen the understanding of the structure and underlying cognitive and social processes. He defined the citation context as the one to three sentences surrounding the citation and in his study used an average of 1.6 sentences surrounding the point of reference. Using 81 full-text papers, co-cited in the organic thin-film transistor domain, he derived a co-citation map and found that sentiments varied within a specialty and were related to cognitive and social factors.

In short, until the early 1990s, semantic CCA mainly relied on manual content analysis over a small sample size and applied survey to shed light on citation motivations and citation functions. However, since the mid-1990s, semantic CCA has been geared towards the application of data mining or natural language processing algorithms to enable semi-automatic analysis of citation contexts.

3. Approaches

Approaches applied in CCA include content analysis as the manual approach and computational linguistics/Natural Language Processing (NLP) as the semi-automatic approach. Content analysis (CA) examines both the syntactic and semantic context of citations to obtain a better understanding of the relationships between citing and cited works. Computational linguistics/Natural Language Processing (NLP) analyzes the citation sentences based on linguistic principles.

3.1 Content Analysis

Traditionally, researchers have employed content analysis (CA) to determine authorship (i.e. identifying personalized linguistic and rhetorical characteristics), examine patterns in documents, and infer psychological or emotional states. In library and information science (LIS) studies, researchers have extended CA to analyze different types of data (e.g. reference interviews, problem statements in published articles, and job advertisements) in both qualitative and quantitative studies. A summary of selected examples of studies in LIS between 1991 and 2005 (White & Marsh, 2006) included the use of CA to identify reasons for selecting initial web search strategies (White & Iivonen, 2001), to develop a thesaurus of image-text relationships (Marsh & White, 2003), and to determine the nature of problem statements in LIS articles (Stansbury, 2002). Similarly, Pettigrew and McKechnie (2001) used a CA codebook of three categories (e.g., Affiliation of First Author, Primary Subject of Article, and Type of Article) to analyze the use of theory in 1,160 articles that appeared in six information science (IS) journals between 1993 and1998.

Still, CA remains not widely applied in citation analysis, despite the fact that the idea of combining bibliometric methods with the full-text analysis for the purposes of content analysis of citations (Cronin, 1984) was put forward and experimented with as early as 1960s (Glenisson, Glänzel, Janssens, & Moor, 2005) when Lipetz identified 29 different citation reasons. In the 1970s and 1980s, a number of researchers (Chubin & Moitra, 1975; Frost, 1979; McCain & Turner, 1989; Moravcsik & Murugesan, 1975; Oppenheim & Renn, 1978; Peritz, 1983; Spiegel-Rösing, 1977) added to this concept by proposing their own schemes to categorize and contextualize citations.

Most CA studies, which examine citation reason and function, employ classification schemes and; over the years, researchers have continually devised schemas, which provide new and different perspectives on

content. Table 1 summarizes the eight prominent schemas detailed as follows: Lipetz's (1965) original study used 29 citation reasons contained in four groups. Chubin and Moitra (1975) proposed a tree hierarchy depicting the cited papers' different levels of contribution (basic, subsidiary, additional, perfunctory) and confirmative or negative nature. Moravcsik and Murugesan (1975) employed four basic binary concepts to categorize citation motivation. The Spiegel-Rösing (1977) schema was very detailed and focused on the contents of the cited work and evaluative use of the citation. Oppenheim & Renn (1978), in one of the largest early CA studies, used seven categories based on their function within the paper (i.e. to support theory, methodology, background). Frost (1979) divided citation usage based on the type of cited text (e.g. other scholar views, primary text, previous scholarship). Peritz (1983) divided citing papers by their roles in the empirical study (e.g. comparative, argumentative, documentary). McCain and Turner (1989) based their classification schema on the location (i.e. introduction, discussion) and scope (central, peripheral) of the citation within the paper.

Recently, as reflected in the work of Zhang, Ding, and Milojević (n.d.), there has been a renewed appreciation for CA's flexible methodology, which allows for both quantitative and qualitative approaches regardless of manual or computer-aid processing. They developed a codebook to annotate citation motivations and better analyze the rich socio-cultural context of citing behavior, which is two-dimensional (citing and cited), bi-modular (syntactic and semantic), and based on the grounded theory (Glaser & Strauss, 1967). Their approach balances specificity and generalizability, while investigating the interaction between individual norms (e.g., personal motivations) and collective norms (e.g. established regulations/conventions in a certain domain) in citing behavior thus continuing the scholarly discussion of content analysis through citation analysis.

Table 1: A comparison of content analysis (CA) classification categories developed for schemas in citation content studies between 1965 and 1989.

	Background Information	Theoretical Framework	Prior Empirical/ Experimental Evidence	Negative Distinction	Methodology Explanation
Lipetz (1965)	▪ non-scientific contribution	▪ scientific contribution	▪ continuity relationship ▪ disposition of contribution		

Chubin & Moitra (1975)	■ affirmative-basic ■ affirmative-additional		■ affirmative-perfunctory ■ affirmative-subsidiary	■ negative-partial ■ negative-total	
Moravcsik & Murugesan (1975)	■ evolutionary or juxtaposition	■ conceptual or organizational	■ perfunctory or organic	■ confirmative or negative	
Spiegel-Rösing (1977)	■ history/state-of-art ■ data (in text) ■ data (in tables)	■ concept definitions ■ new interpretation	■ point of departure ■ data (comparative) ■ further reading ■ substantiated study	■ positive evaluation ■ negative evaluation ■ disproved prior study	■ method
Oppenheim & Renn (1978)	■ historical background ■ data (not comparative)	■ theoretical equation	■ relevant work ■ data (comparative)	■ theory/ method not applicable	■ methodology
Frost (1979)	■ factual evidence	■ views of other scholars	■ primary text ■ further reading ■ previous scholarship		
Peritz (1983)	■ setting stage ■ background ■ documentary		■ comparative ■ argumentative		■ methodology
McCain & Turner (1989)	■ introduction-central ■ introduction-peripheral		■ results & discussion-central ■ results & discussion-peripheral		■ methods-central

3.2 Computational Linguistics/Natural Language Processing

Identifying key concepts within a citation context is a complex problem as many citation contexts are hard to identify, especially using linguistic markers. Computational linguistic techniques using natural language processing has proven useful in key concept extraction, even though scientific citing behavior can vary from field to field and from author to author.

Citation Context: Window Size

Citation context analyzes the textual information located in a window near citation references. Bradshaw (2003) used a fixed window size of 100 words (50 words on either side of the citation) to extract citation context using CiteSeer and by following the "Context" link from the "Document Details" page representing each document. O'Connor (1982) applied a set of hand-crafted sentence-based rules to select the citation context likely to convey information about a particular cited paper over a collection of chemistry journal articles. He concluded that, while helpful for retrieval, it was difficult to identify proper citation context semi-automatically, which requires human intervention and is domain specific. The amount of text referring to a citation can vary dramatically and textually close citations can interact with each other. Ritchie, Teufel and Robertson (2006) demonstrated this variation of citation context and found that almost any window size would result in overlapping windows, which could be attributed to the wrong citation. Using a fixed window size of 50 terms on each side of the reference, similar to Bradshaw (2003), the authors discussed potential computational linguistic techniques such as, some form of text

segmentation, full-blown discourse analysis, or simple sentence boundary detection, as well as, altering the window size to more accurately locate citation contexts.

Ritchie, Robertson and Teufel (2008) tested different citation contexts with the goal of improving information retrieval. As part of their study, they defined citation context in the following nine different ways: *none* indicates no citation context, *1sent* uses only the citation sentence, *3sent* contains the citation sentence plus one sentence immediately to the left and right, *1sentupto* contains the citation sentence and truncated at the next citation to the left and right, *3sentupto* contains the 3sent context and truncated at the next citation to the left and right, *win50* indicates a window of up to 50 words on each side of the citation, *win75* indicates a window of up to 75 words on each side of the citation, *win100* indicates a window of up to 100 words on each side of the citation, and *full* contains the entire citing paper. They assumed the text conformed to grammatical and rhetorical conventions, and that words likely to describe the cited paper occur close to the citation, while words further away were less likely to describe the cited paper. As such, the sentence containing the citation would be a good approximation of the citation's descriptive terms. However, they found that longer citation contexts resulted in greater retrieval effectiveness, *3sent* was more effective than *1sent*, but that truncated versions usually ranked lower, with *1sentupto* below *1sent*. They, therefore, concluded that using neighbor citations to delimit a citation's context was not helpful. While the window context *win50* usually ranked lowest of all the window contexts, increasing the context length did not guarantee better identification of citation contexts since effectiveness decreased again by the time the entire citing paper was taken as the citation context. Similarly, a comparison between the relative effectiveness of the sentence-based and window contexts shows that sentence-based contexts were more effective than windows of equivalent length.

Paper Sections

In her dissertation, Teufel (2000) noted that the diversity of writing styles from different disciplines would derive different paper sections. She found that social scientists tend to utilize unstructured text without standardized section headings, and that 74% of all headers were not prototypical. Specifically, 32% of all papers contained no explicitly mentioned conclusion section and only 9% of the computational linguistics corpus had a background or literature review section. Conversely, section structure in the medical corpus (Cardiology) was very homogeneous with the typical introduction, method, result, and discussion sections existing in almost all papers. Writing styles also varied between regions. For example, German-Polish tradition keeps the results hidden until the end of the work in order to retain readers' curiosity, while English texts provide a results summary in the abstract. As many full-text papers are now available in PDF or HTML format, some researchers have sought to find ways to utilize such formats to identify paper sections. Ding, Liu et al.(2013) used regular expression rules to capture paper sections from the HTML version of Journal of the American Society for Information Science and Technology (JASIST) articles and achieved high accuracy. Ramakrishnan, et al. (2012) developed a layout-aware PDF text extraction system to enable accurate extraction of sections or bodies of text from PDF versions of research articles.

4. Applications

Content-based citation analysis can be divided into the five application categories which are examined in-depth in the following subsections. First, citation motivation classification categorizes citation motivation based on manually annotated training examples or linguistic rules. Second, citation summarization summarizes sets of documents using citation features. Third, information retrieval seeks to enhance

information retrieval performance based on citation context. Fourth, citation recommendation/prediction suggests references for articles or books. Finally, knowledge graph mining identifies the extent to which candidate concepts, from the citation context, can form a conceptual network to enable knowledge discovery.

4.1 Citation motivation classification

Citation count remains the dominant measure of article impact (Borgman, 1990; Ziman, 1968). However, this measurement is often too simplistic to reflect the diverse impact of different types of citations. As a result, many approaches have been proposed and applied to better detect the nuances between papers and their citations and thus build more precise impact measures (Angrosh, Cranefield, & Stanger, 2010; Athar, 2011; Pham & Hoffmann, 2003; Teufel et al., 2006a, 2006b). Garzone (1997) treated the citation classification as a task of sentence categorization using cue words in citations, along with lexical and grammar rules, to break down citation contexts into 35 pre-defined categories. He further divided these categories into the following types: negational, affirmational, assumptive, tentative, methodological, interpretational/developmental, future research, use of conceptual material, contrastive, and reader alert.

The automatic detection of citation function has also been well studied by Teufel (2000) in her doctoral dissertation where she formalized the rhetorical multi-classification task — *Argumentative Zoning* — by labeling sentences as *Own, Other, Background, Textual, Aim, Basis,* and *Contrast* based on its role in the author's arguments. Table 2 displays her final annotation scheme. In her later work, Teufel used 360 conference papers to investigate the problem of classification of citation sentences based on their functional roles (Teufel et al., 2006a, 2006b) and introduced a 12-category citation annotation scheme. She and her colleagues then converted that to a positive/negative/neutral scheme, thus defining a relationship between the binary sentiment classification and the citation function classification.

Table 2: Teufel's Final Annotation Scheme (2000)

Categories	Specification
BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work method, results, future work.
AIM	Specific research goal
TEXTUAL	Textual section structure
CONTRAST	Contrast, comparison, weakness of other solution
BASIS	Other work provides basis for own work

More recently, Angosh et al. (2010) tackled a similar problem of annotation scheme by annotating every sentence in the related work section of an article, including sentences referring to the background and those about the authors' own research. Unlike other works, they viewed the task as sequential labeling and used Conditional Random Fields (CRF) based on the assumption that authors follow a sequential rhetorical pattern while drawing upon related work. Both Teufel and Angosh used heuristically created schemas, which suffered from the problems of multi-classification class imbalance, annotation difficulty,

and limited labeled data. To simplify the number of citation categories for functional classification, Athar (2011) carried out sentiment analysis by reducing the categories into just three classes: positive, negative, and neutral. Their results proved more robust than previous classifications with more classes. Based on an earlier schema (Moravcsik & Murugesan, 1975), Dong and Schäfer (2011) defined a citation classification schema with four types: background, fundamental idea, technical basis, and comparison. Their methodology involved automatic classification through supervised learning classifiers using the textual, physical, and syntactic feature sets. Their results confirmed that the feature set, with the POS tags and added syntactic patterns, was most effective.

There are two major methods used to apply these schema annotations to the citation text. The first stream of research applies a rule-based strategy based on pre-defined cue-words or phrases set in a decision tree classification to classify extracted citations (Garzone, 1997; Nanba, Kando, & Okumura, 2000; Pham & Hoffmann, 2003). The second stream of research employs machine learning techniques to build different classifiers, including IBk (K-NN), Support Vector Machine (SVM), and CRF (Angrosh et al., 2010; Athar, 2011; Teufel et al., 2006a). However, they did not simply use the machine learning-based classifiers alone, but rather integrated expert knowledge in the form of either lexicon (scientific terms) or phrases (cues). Siddharthan and Teufel (2007) applied k-NN to generate intermediate results for the functional classification. They also tested other classifiers including, Naive Bayes, Hidden Naive Bayes, iBk (k-NN), J48 (decision tree), and STACKING (assembling NB and J48). Table 3 summarizes these citation classification studies based on their difference in schema categories, features, and classifiers. Generally, citation classification is similar to sentiment classification, but more complex as the categorization of functional citations remains debatable and usually contains six or even dozen of categories.

Table 3: A comparison of citation motivation annotation schemas, including schema categories, features, and classifiers, developed for citation content studies between 2006 and 2011.

	Schema Categories	Schema Features	Schema Classifiers
Teufel et al. 2006a	▪ Background/ Introduction ▪ Citation Sentences ▪ Descriptive Sentences ▪ Research Gap ▪ Alternate Approach ▪ Current Work	▪ Cue phrases ▪ Verb tense/voice ▪ Modality ▪ Location ▪ (paper/paragraph)	▪ IBk (k-NN)
Teufel et al. 2006b	▪ Weakness ▪ Contrast ▪ Positive ▪ Neutral	▪ Human annotation	
Angrosh et al. 2010	▪ Citation Weakness ▪ Comparison and Contrast ▪ Citation Positive Sentiment ▪ Sentences Neutral Description	▪ Generalization terms (Lexicon) ▪ (Prev.) Sentence has citations	▪ CRF
Athar 2011	▪ Positive ▪ Negative ▪ Neutral	▪ 1-3 grams ▪ Scientific lexicon ▪ POS-tag ▪ Contextual Polarity ▪ Dependency Structure ▪ Sentence Splitting (removing) ▪ Negation	▪ SVM
Dong & Schäfer 2011	▪ Background ▪ Fundamental idea ▪ Technical basis	▪ Cue words ▪ Boolean and weight ▪ POS-tag	▪ SMO ▪ BayesNet, ▪ NaiveBayes

	▪ Comparison	▪ Location ▪ Popularity ▪ Density/Avg Dens

4.2 Citation summarization

As the amount of information has grown in recent years, many researchers have begun developing two types of automated document summarization: key phrase extraction and sentence summarization. Document summarization using key phrase extraction selects the words or phrases from document tags, while sentence summarization selects the sentence to produce short summary paragraphs about a document. Numerous summarization studies (Abu-Jbara & Radev, 2011; Elkiss et al., 2008; Kupiec, Pedersen, & Chen, 1995; Mei & Zhai, 2008; Mohammad et al., 2009; Nanba & Okumura, 2004; Qazvinian, Hassanabadi, & Halavati, 2008; Qazvinian & Radev, 2008, 2010; Teufel & Moens, 2002) have been applied to scientific literature. These studies have sought to identify the minimum size of text necessary to provide the most significant (impact-based), most original (no one addressed yet), and most concise (diversified without redundancy nor lose) information about a paper.

Teufel and Moens' (2002) study was similar to Kupiec et al. (1995), but instead of evaluating document summarization based on domain expert summaries, Teufel and Moens prepared the summaries themselves. Again, similar to Kupiec et al., Teufel and Moen relied on both supervised learning and human-selected extraction for sentence extraction, but their training and evaluation were based on good candidate sentences, and extraction was based on human judgment. Teufel and Moens' results were not significantly different from that of Kupiec et al., which implied that any improvement should be made to extraction methods rather than to the training set.

While earlier studies (Nanba et al., 2000; Nanba & Okumura, 1999) analyzed citation sentences using pre-defined phrase-based rules to build survey generation tools, Qazvinian and Radev (2008) treated citation sentences as resources for fact summarization. They selected a subset of citation sentences to form a summary based on defined criteria and assuming that the ideal summary is composed of the most important and diversified facts. To find such facts, they first clustered the citation sentences and then applied a network-based ranking algorithm within each cluster. Elkiss, et al. (2008) similarly studied citation summaries generated from research papers in PubMed. Deeming citation summary information as that which is important to peers, they showed that citation summaries both overlap and differ from the paper abstracts since each focused on different aspects of the paper.

Mei and Zhai (2008), casting the problem as a retrieval task, proposed a language model-based summarization method. Regarding each candidate sentence in the summary as a document capable of being retrieved, they constructed a virtual impact query. A major contribution of their study was the use of different citation weights based on authority and prestige. Qazvinian and Radev (2010) also used a language model to tackle the problem of automatic key phrase extraction and sentence selection. They used point-wise divergence to measure how randomly a phrase can be generated with respect to its unique words and then set a threshold for key phrase generation. With the goal of picking sentences which include the most important and non-redundant key phrases, they approximated the optimization by greedily adding new sentences into a current solution set. Mohammad, et al. (2009) applied similar techniques while automatically generating a scientific survey for multiple documents.

Other researchers (Abu-Jbara & Radev, 2011; Qazvinian et al., 2008) have proposed using LexRank (a network-based ranking algorithm equivalent to PageRank) to identify the most salient sentences within clusters. LexRank first summarizes multi-documents and builds a cosine similarity graph of all the

candidate sentences. Then it finds the most central sentences by performing a random walk on the graph. LexRank sets each citation sentence as a node and their similarity as the edge weight. From the nature of stationary distribution of Markov Chain, the most central papers are selected based on the main facts of the corresponding cluster (e.g., representative sentences). Multiple measures for setting the edge weight can be used for reordering the sentences. In fact, the supervised method is often more expensive than the unsupervised method, as it requires prior learning based on sufficient training data. As a result, in recent studies, as shown in Table 4, unsupervised learning has been more frequently used.

Table 4: Methods and Feature differences in citation summarization studies

	Methods	Features
Kupiec et al. (1995)	▪ Supervised learning	▪ Human-selected extraction: professional abstractors
Teufel & Moens 2002	▪ Supervised learning	▪ Human-selected extraction: authors
Qazvinian & Radev 2008	▪ Unsupervised learning (clustering)	▪ LexRank ▪ C-LexRank ▪ C-RR
Elkiss et al. 2008		▪ A lexical similarity metric *Self cohesion*
Qazvinian et al. 2010	▪ Unsupervised learning (clustering)	▪ LexRank ▪ C-LexRank ▪ C-RR ▪ MMR ▪ Key phrase extraction ▪ N-gram
Abu-Jbara & Radev 2011	▪ Unsupervised learning (clustering)	▪ LexRank ▪ Variations: remove sentence filtering or classification or clustering component
Mei & Zhai 2008	▪ A language model-based summarization method	▪ The use of different citation weights based on authority and prestige

4.3 Retrieval

Citation information, in the form of citation counts, has shown marginal performance improvement in information retrieval. Meij and Rijke (2007) used citation counts to identify the prior probability of a document's appropriateness in the language model retrieval framework and Fujii (2007) used PageRank to calculate citation impact for improved patent search. Other studies have sought to enhance information retrieval performance using citation information (e.g., citation sentence, and citation context) by finding the index terms of a paper in the citation context (O'Connor, 1982).

Bradshaw (2003) proposed an indexing technique, Reference Directed Indexing (RDI), which combined measures of relevance and significance in a single retrieval metric based on a comparison of the terms authors used in reference to documents. While citation frequency measures are useful in determining the relative importance of documents, it has been difficult to determine the relevance of such documents to a given query. Leveraging the fact that sufficiently useful documents are cited by multiple authors and using the terms appearing in the citation context, RDI can establish the relevance of a reference to a given query term. Repeated references to a document provide a means of comparing the words of many

references. If several authors use the same words in reference to a document, RDI views these words as good index terms for that document. He, et al. (2010) added keywords and other citation metadata (e.g., title, cited articles) to represent the citing document and applied the k-means clustering algorithm to group the retrieved documents into different categories thereby boosting retrieval performance.

These studies combined the citation context with only partial information about the cited paper (e.g., citation metadata). Ritchie, et al. (2008), however, combined citation contexts with the full-text of the cited paper. They added the context of citing papers into the cited papers to improve the retrieval performance for cited articles. Using fixed window words, truncated words, and the full paper to index cited papers, they found that using longer citation contexts could improve the retrieval performance, but that results worsened when the full paper was used to index cited articles. In addition, their research indicated that weighting citation terms higher than document terms generally improved retrieval effectiveness.

Qazvinian and Radev (2010) also used context sentences surrounding citations determine relevance and increase retrieval performance. They aimed to solve the problem of retrieval for context sentences with respect to the given query ("reference-paper" pair), and addressed this task by considering the nature of connections between paper and reference. They used the Markov Random Field (MRF) model as a collective classifier to label candidate sentence as "relevant" or "non-relevant."

4.4 Citation Recommendation/Prediction

Citation context is considerably useful for constructing a citation recommendation system to find related work. Nallapati, et al. (2008) exemplified this approach by proposing Pairwise-Link-LDA, which models the existence of a link between every pair of documents based on words generated from a topic-word distribution. Kataria, et al. (2011) extended this model to cite-PLSA-LDA, by associating terms in the citation contexts to the cited documents, and generating topic-citation multinomial distributions in the citing paper. Similarly, Tang and Zhang (2009) proposed a two-layer Restricted Boltzmann Machine model (RBM-CS), which could discover topic distributions of paper content and citation relationship simultaneously. In this way, they provided a discriminative approach to topic-based citation recommendation.

Based on CiteSeerX, He, et al. (2010) proposed a context-aware citation recommendation system to recommend a possible list of bibliographic records for a given manuscript. Their core idea was to design a novel non-parametric probabilistic model to measure the context-based relevance between a citation context and a document. Later, He, et al.(2011) built a citation context based system, which sought to identify locations within a query manuscript where citations were needed. They proposed four models for finding citation contexts: language models, contextual similarity, topic relevant, and dependency feature model. Huang, et al. (2012) developed a translation method to convert research papers into references. They represented research papers using both the descriptive language (words appearing in the citation sentences) and the reference language (features of the references). Their citation recommendation system, thus transformed citation context into a representation for the cited papers.

Citation context has also been used in citation prediction, which usually focuses on predicting links between networks of documents, and on predicting words within them. For example, Chang and Blei (2009) developed the relational topic model (RTM), which modeled data composed of documents (i.e., collections of words and links among words). For each pair of documents, the RTM modeled their link as "a binary random variable that is conditioned on their contents" (p. 81). Focusing on the potential correlation between topic similarity and community closeness, Liu, Niculescu-Mizil and Gryc (2009)

developed the Topic-Link LDA model to jointly model topics and author community. Using this model, the similarity between topic mixtures via citation links could help predict the similarity between community mixtures, or vice versa. Dietz, Bickel, and Scheffer (2007) built a probabilistic topic model that includes citation content variables such as topic mixture of the topical atmosphere of a cited publication and characteristic word distribution for each topic. Using this model, the strength of influence of citations against manually rated citations can be predicted.

4.5 Knowledge Graph Mining

Many researchers have attempted to extract important concepts from citation context to form conceptual networks (Rees-Potter, 1989; Schneider & Borlund, 2004; Schneider, 2006). Small (1978) argued that citations reflect the author's commentary on the cited work through the process of making symbols and creating meaning. The content of the citation context, as a symbol of concepts and methods, can therefore be mined for meaning. Ding, Song, et al. (2013) proposed entitymetrics to extend bibliometric methods by measuring the impact of knowledge entities in scholarly communication. They defined knowledge entities as those entities which act as carriers of knowledge in scientific articles, such as, keywords, topics, subject categories, datasets, key methods, key theories, and domain entities (e.g., biological entities: genes, drugs, and diseases) (see Figure 3). They tested the usefulness of this approach by analyzing the knowledge entities in PubMed Central documents related to the drug Metformin. They formed a biological entity citation network and analyzed the features of the network and node centralities (see Figure 4). Comparing their results with the manually curated Comparative Toxicogenomics Database (CTD) demonstrated the usefulness of entitymetrics in detecting the most outstanding biological entities related to the drug Metformin.

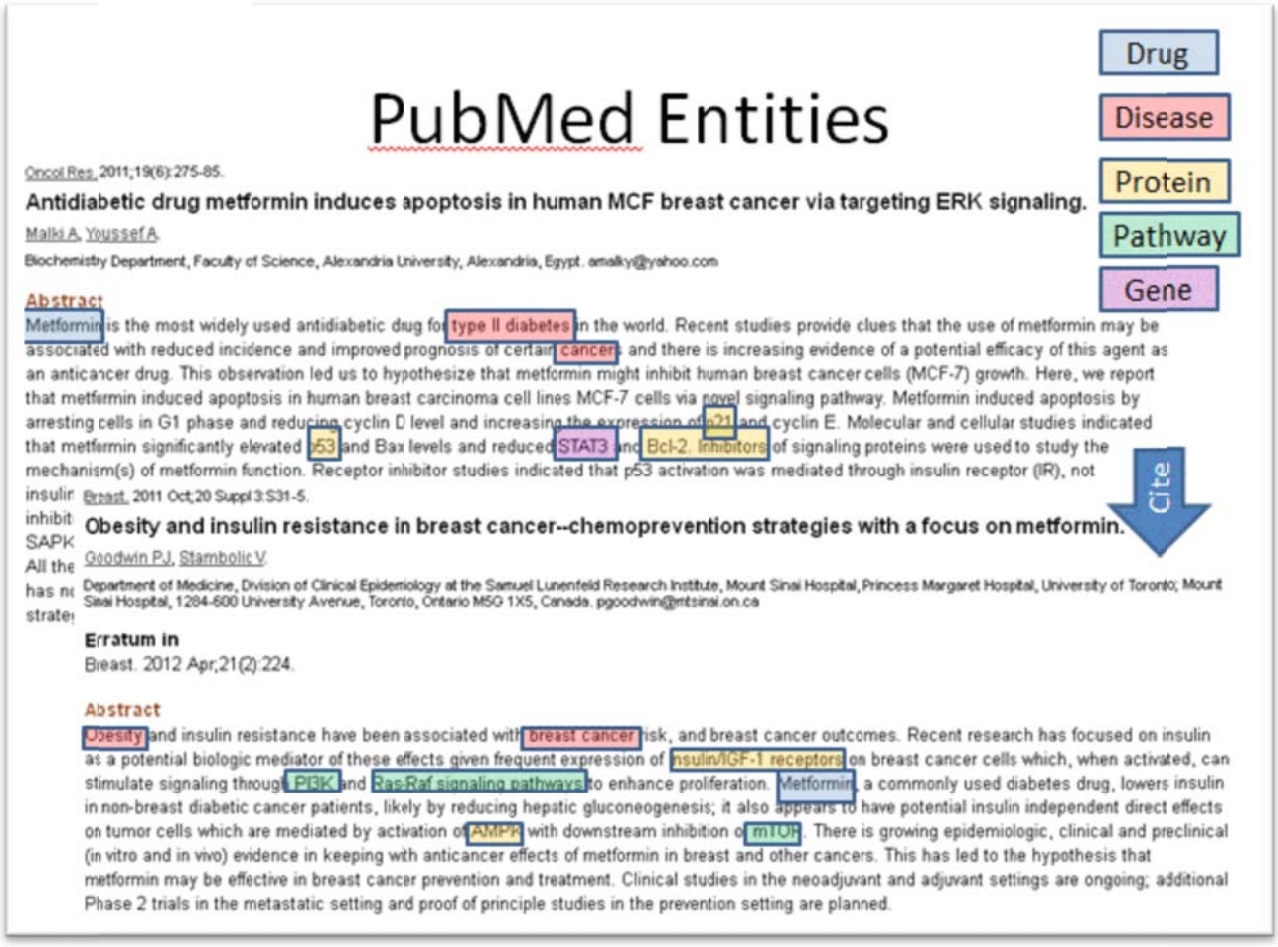

Figure 3: Biological entities highlighted in two PubMed articles which one cites the other.

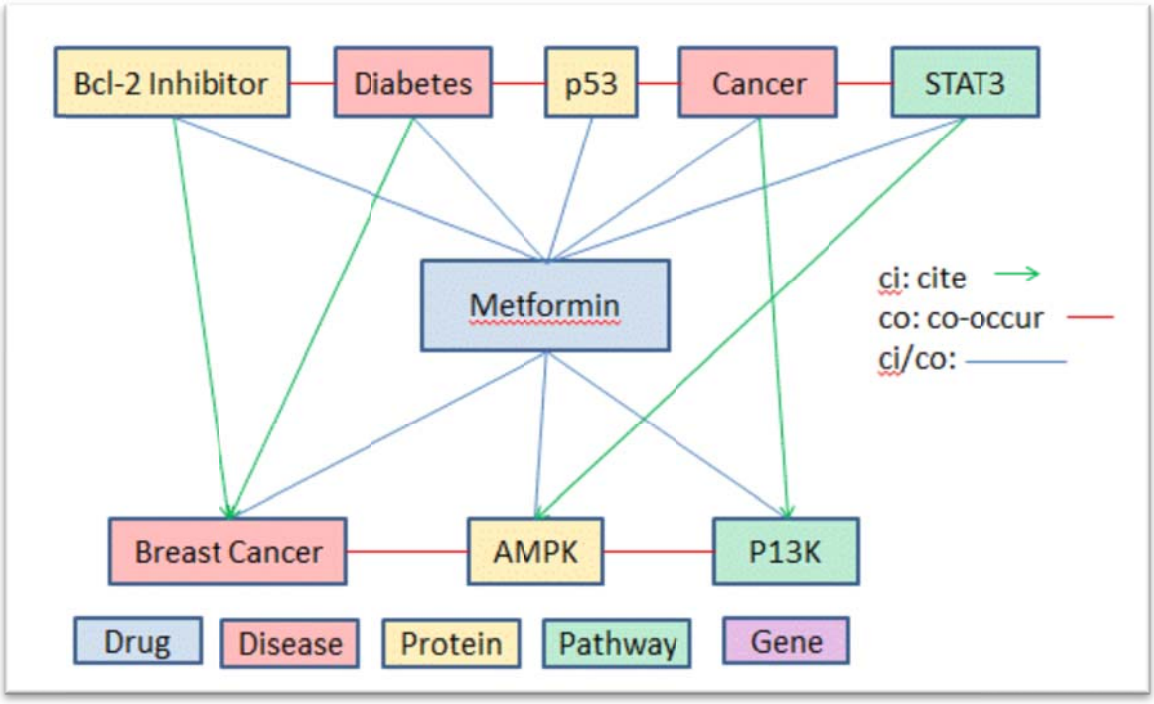

Figure 4: The heterogeneous entity graph based on entity citing, co-citing and co-occurring relationships

Similarly, Jenssen, et al. (2001) created a gene-gene co-citation network for 13,712 human genes by extracting gene names from 10 million Medline articles. They interlinked this co-citation network with Medical Subject Headings (MeSH) terms and the gene ontology (GO) database to validate that co-citation associations between gene entities reflect meaningful biological relationships. Arnold and Cohen (2009) used citation networks to help identify genes in academic articles. After analyzing which genes an author had written about, with whom he/she coauthored, and which articles or authors he/she usually cited they predicted which genes the author would write about in the future. Through this, they demonstrated that scholarly communication and network analysis could provide better link prediction than relying solely on traditional biological information.

Schlitt, et al. (2003) proposed a method to identify functionally related genes based on comparison of neighborhoods in gene networks. They proposed that if two nodes had a large overlap between their neighborhoods, and then they were more likely to link together. The neighborhood was measured using protein-protein interaction data, protein complexes, and a literature network constructed based on co-occurrences of gene names in abstracts of scientific articles. Using this method, they identified 816 functional relationships between 159 genes and assigned biological process annotation to seven previously uncharacterized genes.

5 Conclusion and Future Directions

Scientists write articles to document their research outputs and publish them to disseminate knowledge. Published scientific articles are tested, evaluated, improved/criticized, and applied in different scientific disciplines to generate more research articles. This loop forms the way modern science accumulates and develops knowledge. Research papers, while not the only form of the scientific output, do authenticate the

accreditation process of how knowledge is developed. Therefore, analyzing the content of scientific articles, especially through citing behavior, can identify how knowledge has evolved. This paper provides a comprehensive understanding of the state of the art of content-based citation analysis (CCA).

The foundation of CCA seeks to explain the "how" and "why" of citation behavior either through syntactic and semantic analysis. Syntactic CCA uses the structure or layout of an article to identify the location and number of citations to investigate the significance of influence. Semantic CCA deepens this analysis by identifying citation motivation based on pre-defined categorizations through manual analysis of citation context. Given that most analysis of this type is based on a limited number of citing articles (e.g., usually less than 100 articles), the challenge remains on how to generalize these findings to other citing articles.

Current approaches to CCA include both this manual approach of content analysis and the semi-automatic approach of natural language processing (NLP). Content analysis has been widely applied in computer-mediated communication to detect communication patterns among different communities. The core part of CCA is the development of a codebook used to annotate citation contexts. To automatize the extraction of key concepts from citation contexts, researchers use NLP, which allows for the analysis of citing behavior. It is however, still challenging to identify the best window size to extract the proper citation context and to detect the correct citing paper sections.

Researchers have applied CCA in various ways to facilitate better management and evaluation of research behavior, summarization and retrieval of information, recommendation and prediction of scholarly communication, and mining and discovery of knowledge. Citation motivation classification uses a rule-based approach, based on pre-defined cue-words, to classify motivations semi-automatically. Following this approach, studies have used citations, together with abstracts or full-text contents, to generate summaries of specific sub-domains. Following the success of PageRank, many began implanting citations contexts and topic features into retrieval algorithms to improve performance. Recommender systems, in demand as seeking good related works is no longer trivial, have used citations to build relationships between authors, topics, articles, and publication venues.

With the launch of Google's knowledge graph, as a part of Google's semantic search initiatives in 2012, and the increasing interest in big data, concept-driven or entity-driven graph mining has surged. While there are many ways to form knowledge graphs, citations and their contexts provide a unique link to connect concepts or entities. Using the large-size publicly available PubMed articles and full-text PubMed central articles to build entity citation graphs or entity co-citation graphs has contributed to the discovery of unknown knowledge (Ding, Song, et al., 2013). How to better integrate these entity citation graphs with other domain related graphs (e.g., with other publicly available databases about genes, drugs, diseases, and side effects) to enable intelligent knowledge discovery is an interesting direction for future research. Similarly, systems to predict or recommend citations, as well as, those using citations to generate summaries all have space to improve. Questions such as, how to identify patterns of contexts used by papers citing a specific article and how to use those patters to predict future citations, or how to leverage citation contexts by giving more weights to important citations in citing articles thus generating better summaries have yet to be answered and will surely provide directions for the future research. Finally we would like to call for the following initiatives:

Challenge for the art of scholarly writing

Along with people contributing data via shopping, purchasing, photographing, tweeting, and blogging, researchers are contributing data through conducting their research experiments and writing their research outputs. The data we are generating is expanding exponentially, especially due to the contribution of social media and usage of the Internet. This is big data, and it will get even bigger. Big data brings us

opportunities, as well as, challenges. Scientific papers are unique in that they allow for the tracing of data from one researcher to another by citation of their work or quotation of their statements. Nowadays, however there are too many articles to read and cite, and as a result, intelligent recommender systems, such as Google Scholar, play a crucial role in recommending articles. Will this, citing articles suggested automatically by recommender systems, challenge the art of scholarly writing? This could become an interesting research topic to explore. On one hand, researchers now depend on retrieval systems or recommender systems to find related works for them. Because of this increase in data, they have lost the capability to browse the whole set of related documents, which was not an issue ten years ago. They must ensure retrieval systems or recommender systems sample the right set of documents to read and cite. On the other hand, researchers can now identify related and high quality papers based on citation frequency or journal impact factors. Of course, broadcasting papers using tweets or blogs can increase an articles' visibility and citations. It is high time we revisit the citing behavior research of researchers in the late 60s and 70s to see whether big data has influenced our way of writing and citing.

Entitymetrics

Currently knowledge is encoded as strings in unstructured scientific literature, which creates a huge hurdle for fast knowledge dissemination and industry transfer. Lifting knowledge out of the unstructured article, in PDF format, will help us connect the dots and be more innovative. Knowledge entities/concepts broadly include keywords, topics, subject categories, datasets, key methods, key theories, and domain entities (e.g., biological entities: genes, drugs, and diseases). These knowledge entities are often used to mine knowledge and can be used ultimately to facilitate knowledge discovery based on their ability to co-occur, cite/being cited, or co-cite/being co-cited. For example, co-author connections in articles can reflect scientific collaboration patterns and gene co-occurrence connections in articles can identify potential association among genes. The overlay of co-author networks with gene co-occurrence networks can portray the entity-oriented scientific collaboration landscapes. Entitymetrics can bring a paradigm shift to bibliometrics by taking knowledge entity as the research unit to enable knowledge discovery (Ding, et al., 2013).

Paradigm shift for scholarly publishing

The future success of scholarly publishing depends on whether we can create an ecosystem of scholarly products to enable immediate knowledge transfer (Byrnes et al., 2013). In the current fast-moving big data era, especially the data-driven or data-intensive sciences, text is no longer the most efficient way to convey scientific information (Mons et al., 2011). In order to create such an ecosystem, we need to extract knowledge units from scientific papers and represent scientific results in a machine-readable format (e.g., RDF (Resource Description Format) from W3C) so that these knowledge units/concepts and claims can be automatically or semi-automatically linked to related ones. In such a way, a huge knowledge base can be formed organically. Of course, further technologies will be developed to advance such processes and ensure quality. For example, the nanopublication initiative is developing provenance and context semantics to support knowledge discovery and connect treasures of implicit information (Mons et al., 2011). This new shift will generate different ways of writing and citing. Authors can cite knowledge units/concepts or scientific claims right after a paper has been accepted and not yet published. One paper can be cited multiple times if it contains several scientific claims or knowledge concepts. An author's citation of one knowledge unit can lead them to cite the linked knowledge units from this huge knowledge base. Therefore, a citation can be more modularized, which means that it does not have to be a paper, it can be a part of a paper (e.g., a scientific claim in the conclusion, or an experimental setting in the methodology part). Scholarly writing will change as well, for example; authors normally provide keywords for their papers; instead, they will be asked to provide several scientific claims and important knowledge units/concepts as part of their submissions. This will turn traditional scholarly communication into a new paradigm for sharing and transferring knowledge units, data, and scientific claims, which enables fast knowledge discovery.

References

Abu-Jbara, A., & Radev, D. R. (2011). Coherent Citation-Based Summarization of Scientific Papers. In *49th Annual Meeting of the Association for Computational Linguistics, June 19-24, Portland, Oregon Proceedings* (pp. 500–509).

Angrosh, M. A., Cranefield, S., & Stanger, N. (2010). Context identification of sentences in related work sections using a condidtional rankdom field: Towards intelligent digital libraries. In *Annual Joint Conference on Digital Libraries, JCDL'10 Proceedings* (pp. 293–302). doi:10.1145/1816123.1816168

Arnold, A., & Cohen, W. W. (2009). Information Extraction as Link Prediction: Using Curated Citation Networks to Improve Gene Detection. In B. Liu, A. Bestavros, D.-Z. Du, & J. Wang (Eds.), *Wireless Algorithms, Systems, and Applications: 4th International Conference, WASA 2009, Boston, MA, USA, August 16-18, 2009. Proceedings* (pp. 541–550). Springer Berlin Heidelberg. doi:10.1007/978-3-642-03417-6_53

Athar, A. (2011). Sentiment analysis of citations suing sentence structure-based features. In *2011 Student Session, ACL-HLT , Portland, Oregon June 19-24, Proceedings* (pp. 81–87). Association for Computational Linguistics.

Becher, T. (1989). *Academic tribes and territories: Intellectural enqiry and the culture of disciplines*. Oxford, UK: Oxford University Press.

Bonzi, S. (1982). Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science*, *33*(4), 208–216. doi:10.1002/asi.4630330404

Borgman, C. L. (1990). *Scholarly communications and bibliometrics*. Newberry Park, CA: Sage Publications, Inc.

Bradshaw, S. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. In T. Koch & I. T. Sølvberg (Eds.), *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003 Trondheim, Norway, August 17-22, 2003 Proceedings* (pp. 499–510). Springer Berlin Heidelberg. doi:10.1007/978-3-540-45175-4_45

Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science and Technology*, *36*(4), 223–229. doi:10.1002/asi.46303360402

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science and Technology*, *40*(4), 284–290. doi:10.1002/(SICI)1097-4571(198907)40:4<284::AID-ASI10>3.0.CO;2-Z

Chang, J., & Blei, D. M. (2009). Relational Topic Models for Document Networks. In *Proceedings of the 12th International Confe-rence on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.* (pp. 81–88).

Chubin, D. E., & Moitra, S. D. (1975). Content Analysis of References: Adjunct or Alternative to Citation Counting? *Social Studies of Science*, *5*(4), 423–441. Retrieved from http://www.jstor.org/stable/284806

Cronin, B. (1981). The need for a theory of citing. *Journal of Documentation*, *37*(1), 16–24. doi:10.1108/eb026703#sthash.rmBNf1qS.dpuf

Cronin, B. (1984). *Citation process: Role and significance of citations in scientific communication*. Taylor Graham.

Cronin, B. (2005). *The Hand of Science* (pp. 1–7). Lanham, Maryland: Scarecrow Press.

Dietz, L., Bickel, S. S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning , ICML '07* (pp. 233–240). doi:10.1145/1273496.1273526

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, *7*(3), 583–592. doi:10.1016/j.joi.2013.03.003

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PLoS ONE*, *8*(8), 1–14. doi:10.1371/ journal.pone.0071416

Dong, C., & Schäfer, U. (2011). Ensemble-style Self-training on Citation Classification. In *5th International Joint Conference on Natural Language Processing, IJCNLP 2011, Proceedings* (pp. 623–631).

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, *59*(1), 51–62. doi:10.1002/asi.20707

Frost, C. O. (1979). The Use of Citations in Literary Research: A Preliminary Classification of Citation Functions. *The Library Quarterly: Information, Community, Policy*, *49*(4), 399–414. Retrieved from http://www.jstor.org/stable/4307148

Fujii, A. (2007). Enhancing patent retrieval by citation analysis. In *30th annual international ACM SIGIR conference on Research and development in information retrieval,SIGIR '07, Proceedings* (pp. 793–794). doi:10.1145/1277741.1277912

Garfield, E. (1964). Can Citation Indexing be Automated? In M. E. Stevens, Vincent E. Giuliano, & L. B. Heilprin (Eds.), *Statistical Association Methods for Merchanized Documentation: Symposium Proceedings Washington 1964* (pp. 189–192). Department of Commerace National Bureau of Standards.

Garfield, E. (1974). The citation index as a subject index. In *Essays of an Information Scientist: Volume 2, 1974-1975* (pp. 62–64). ISI Press.

Garzone, M. A. (1997). *Automated classification of citations using linguistic semantic grammars*. The University of Western Ontario.

Gilbert, G. N. (1977). Referencing as Persuasion. *Social Studies of Science*, *7*(1), 113–122. Retrieved from http://www.jstor.org/stable/284636

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New Jersey, USA: Transaction Publishers.

Glenisson, P., Glänzel, W., Janssens, F., & Moor, B. De. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, *41*(6), 1548–1572. doi:10.1016/j.ipm.2005.03.021

He, Q., Kifer, D., Pei, J., Mitra, P., & Giles, C. L. (2011). Citation recommendation without author supervision. In *Proceedings of the fourth ACM international conference on Web search and data mining , WSDM '11* (pp. 755–764). doi:10.1145/1935826.1935926

He, Q., Pei, J., Kifer, D., Mitra, P., & Giles, C. L. (2010). Context-aware citation recommendation. In *19th international conference on World Wide Web, WWW '10, Proceedings* (pp. 421–430). doi:10.1145/1772690.1772734

Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science and Technology*, *29*(6), 308–310. doi:10.1002/asi.4630290608

Hodges, T. L. (1972). *Citation indexing: Its potential for bibliographical control*. University of California, Berkeley.

Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, C. L., & Rokach, L. (2012). Recommending citations: translating papers into references. In *21st ACM international conference on Information and knowledge management , CIKM'12, Proceedings* (pp. 1910–1914). doi:10.1145/2396761.2398542

Jenssen, T.-K., Laegreid, A., Komorowski, J., & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, *28*, 21–28. doi:10.1038/ng0501-21

Kataria, S., Mitra, P., Caragea, C., & Giles, C. L. (2011). Context sensitive topic models for author influence in document networks. In *Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three, IJCAI '11, Proceedings* (pp. 2274–2280). doi:10.5591/978-1-57735-516-8/IJCAI11-379

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In *18th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '95, Proceedings* (pp. 68–73). doi:10.1145/215206.215333

Lipetz, B.-A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *Journal of the American Society for Information Science and Technology*, *16*(2), 81–90. doi:10.1002/asi.5090160207

Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009). Topic-link LDA: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning , ICML '09* (pp. 665–672). doi:10.1145/1553374.1553460

Maričić, S., Spaventi, J., Pavičić, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the American Society for Information Science and Technology*, *49*(6), 530–540. doi:10.1002/(SICI)1097-4571(19980501)49:6<530::AID-ASI5>3.0.CO;2-8

Marsh, E. E., & White, M. D. (2003). A taxonomy of relationships between images and text. *Journal of Documentation*, *59*(6), 647–672. doi:10.1108/00220410310506303

McCain, K. W., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics*, *17*(1-2), 127–163. doi:10.1007/BF02017729

Mei, Q., & Zhai, C. (2008). Generating impact-based summaries for scientific literature. In *46th Annual Meeting of the Association for Computational Linguistics (ACL) with the Human Language Technology Conference (HLT) of the North American Chapter of the ACL, June 15-20, Columbus, OH Proceedings* (pp. 816–824).

Meij, E., & Rijke, M. de. (2007). Using prior information derived from citations in literature search. In *RIAO '07 Large Scale Semantic Access to Content (Text, Image, Video, and Sound) Proceedings* (pp. 665–670).

Merton, R. K. (1957). Priorities in Scientific Discovery: A Chapter in the Sociology of Science. *American Sociological Review*, *22*(6), 635–659. Retrieved from http://www.jstor.org/stable/2089193

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., … Zajic, D. (2009). Using citations to generate surveys of scientific paradigms. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics Proceedings* (pp. 584–592).

Moravcsik, M. J., & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, *5*, 86–92. doi:10.1177/030631277500500106

Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In *14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08, Proceedings* (pp. 542–550). doi:10.1145/1401890.1401957

Nanba, H., Kando, N., & Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, *11*(1), 117–134. doi:10.7152/acro.v11i1.12774

Nanba, H., & Okumura, M. (1999). Towards multi-paper summarization reference information. In *16th international joint conference on Artificial intelligence, IJCAI'99, Proceedings* (pp. 926–931).

Nanba, H., & Okumura, M. (2004). Comparison of some automatic and manual methods for summary evaluation based on the text summarization challenge 2. In *LREC: 4th International Conferenc n Language Resources and Evaluation, May 26-28, Lisbon, Portugal Prceedings*.

O'Connor, J. (1982). Citing statements: Computer recognition and use to improve retrieval. *Information Processing & Management*, *18*(3), 125–131. doi:10.1016/0306-4573(82)90036-X

Oppenheim, C., & Renn, S. P. (1978). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science and Technology*, *29*(5), 225–231. doi:10.1002/asi.4630290504

Peritz, B. C. (1983). A classification of citation roles for the social sciences and related fields. *Scientometrics*, *5*(5), 303–312. doi:10.1007/BF02147226

Pettigrew, K. E., & McKechnie, L. E. F. (2001). The use of theory in information science research. *Journal of the American Society for Information Science and Technology*, *52*(1), 62–73. doi:10.1002/1532-2890(2000)52:1<62::AID-ASI1061>3.0.CO;2-J

Pham, S. B., & Hoffmann, A. (2003). A new approach for scientific citation classification using cue phrases. In T. D. Gedeon & L. C. C. Fung (Eds.), *AI 2003 Advances in artificial intelligence: 16th Australian Conference on AI, Perth, Australia, December 3-5, 2003 Proceedings* (pp. 759–771). Springer Berlin Heidelberg. doi:10.1007/978-3-540-24581-0_65

Qazvinian, V., Hassanabadi, L. S., & Halavati, R. (2008). Summarizing text with a genetic algorithm-based sentence extraction. *International Journal of Knowledge Management Studies*, *2*(4), 426–444. doi:10.1504/IJKMS.2008.01975

Qazvinian, V., & Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *22nd International Conference on Computational Linguistics, COLING '08, Proceedings* (pp. 689–696). Stroudsburg, PA, USA: Association for Computational Linguistics.

Qazvinian, V., & Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In *48th annual meeting of the association for computational linguistics, July, 11-16, Uppsala, Sweden Proceedings* (pp. 555–564).

Ramakrishnan, C., Patnia, A., Hovy, E., & Burns, G. A. (2012). Layout-aware text extraction from full-text PDF of scientific articles. *Source Code for Biology and Medicine*, *7*(7), 1–10. doi:10.1186/1751-0473-7-7

Rees-Potter, L. K. (1989). Dynamic thesaural systems: A bibliometric study of terminological and conceptual change in sociology and economics with application to the design of dynamic thesaural systems. *Information Processing & Management*, *25*(6), 677–689. doi:10.1016/0306-4573(89)90101-5

Ritchie, A., Robertson, S., & Teufel, S. (2008). Comparing citation contexts for information retrieval. In *17th ACM conference on Information and knowledge management Proceedings* (pp. 213–222). doi:10.1145/1458082.1458113

Ritchie, A., Teufel, S., & Robertson, S. (2006). How to find better index terms through citations. In *Workshop on How Can Computational Linguistics Improve Information Retrieval? CLIIR '06 Proceedings* (pp. 25–32). Stroudsburg, PA, USA: Association for Computational Linguistics.

Schlitt, T., Palin, K., Rung, J., Dietmann, S., Lappe, M., Ukkonen, E., & Brazma, A. (2003). From Gene Networks to Gene Function. *Genome Research*, *13*, 2568–2576. doi:10.1101/gr.1111403

Schneider, J. W. (2006). Concept symbols revisited: Naming clusters by parsing and filtering of noun phrases from citation contexts of concept symbols. *Scientometrics*, *68*(3), 573–593. doi:10.1007/s11192-006-0131-z

Schneider, J. W., & Borlund, P. (2004). Introduction to bibliometrics for construction and maintenance of thesauri: Methodical considerations. *Journal of Documentation*, *60*(5), 524–549. doi:10.1108/00220410410560609

Shadish, W. R., Tolliver, D., Gray, M., & Gupta, S. K. Sen. (1995). Author Judgements about Works They Cite: Three Studies from Psychology Journals. *Social Studies of Science*, *25*(3), 477–498. doi:10.1177/030631295025003003

Siddharthan, A., & Teufel, S. (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. In *NAACL HLT 2007 Proceedings* (pp. 316–323).

Small, H. G. (1978). Cited Documents as Concept Symbols. *Social Studies of Science*, *8*(3), 327–340. doi:10.1177/030631277800800305

Small, H. G. (1982). Citation context analysis. In B. Dervin & M. J. Voigt (Eds.), *Progress in Communication Sciences (Vol.3)* (pp. 287–310). Norwood, NJ, USA: Ablex.

Small, H. G. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, *87*(2), 373–388. doi:10.1007/s11192-011-0349-2

Snizek, W. E., Oehler, K., & Mullins, N. C. (1991). Textual and nontextual characteristics of scientific papers: Neglected science indicators. *Scientometrics*, *20*(1), 25–35. doi:10.1007/BF02018141

Spiegel-Rösing, I. (1977). Science Studies: Bibliometric and Content Analysis. *Social Studies of Science*, *7*(1), 97–113. Retrieved from http://www.jstor.org/stable/28463

Stansbury, M. C. (2002). Problem statements in seven LIS journals: An application of the Hernon/Metoyer-Duran attributes. *Library & Inforamtion Science Research*, *24*(2), 157–168. doi:10.1016/S0740-8188(02)00110-X

Suppe, F. (1998). The Structure of a Scientific Paper. *Philosophy of Science*, *65*(3), 381–405. Retrieved from http://www.jstor.org/stable/188275

Tang, J., & Zhang, J. (2009). A Discriminative Approach to Topic-Based Citation Recommendation. In T. Theeramunkong, B. Kijsirikul, N. Cercone, & T.-B. Ho (Eds.), *Advances in Knowledge Discovery and Data Mining:13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings* (pp. 572–579). Springer Berlin Heidelberg. doi:10.1007/978-3-642-01307-2_55

Teufel, S. (2000). *Argumentative zoning: Information extraction from scientific text*. University of Edinburgh.

Teufel, S., & Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, *28*(4), 409–445. doi:10.1162/089120102762671936

Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). An annotation scheme for citation function. In *7th SIGdial Workshop on Discourse and Dialogue, sigDAIL'09, Proceedings* (pp. 80–87). Association for Computational Linguistics.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006b). Automatic classification of citation function. In *7th SIGdial Workshop on Discourse and Dialogue, sigDAIL'09, Proceedings* (pp. 80–87). Stroudsburg, PA, USA: Association for Computational Linguistics.

Vinkler, P. (1987). A quasi-quantitative citation model. *Scientometrics*, *12*(1-2), 47–72. doi:10.1007/BF02016689

Voos, H., & Dagaev, K. S. (1976). Are All Citations Equal? Or, Did We Op. Cit. Your Idem? *Journal of Academic Librarianship*, *1*(6), 19–21.

White, M. D., & Iivonen, M. (2001). Questions as a factor in Web search strategy. *Information Processing & Management*, *37*(5), 721–740. doi:10.1016/S0306-4573(00)00043-1

White, M. D., & Marsh, E. E. (2006). Content Analysis: A Flexible Methodology. *Library Trends*, *55*(1), 22–4. doi:10.1353/lib.2006.0053

Zhang, G., Ding, Y., & Milojević, S. (n.d.). Citation content analysis (CCA): A method for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*.

Ziman, J. M. (1968). *Public knowledge: An essay concerning the social dimension of science*. Cambridge, UK: Cambridge University Press.