Exploring Characteristics of Highly Cited Authors according to Citation Location and Content

Juyoung An, Namhee Kim
Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seoul,
South Korea. E-mail: {anjy, nakim}@yonsei.ac.kr

Min-Yen Kan, Muthu Kumar Chandrasekaran
NUS School of Computing, National University of Singapore, AS6 05-12, 13 Computing
Drive, Singapore. E-mail: {kanmy, muthu.chandra}@comp.nus.edu.sg

Min Song
Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seoul,
South Korea. E-mail: min.song@yonsei.ac.kr

Big Science and cross-disciplinary collaborations have reshaped the intellectual structure of research areas. A number of works have tried to uncover this hidden intellectual structure by analysing citation contexts. However, none of them analyze by document logical structures such as *sections*. The two major goals of this study are to find characteristics of authors who are highly-cited sectionwise and to identify the differences in section-wise author networks. This study uses 29,158 of research articles culled from the ACL Anthology, which hosts articles on computational linguistics and natural language processing. We find that the distribution of citations across sections is skewed and that a different set of highly-cited authors share distinct academic characteristics, according to their citation locations. Furthermore, the author networks based on citation context similarity reveal that the intellectual structure of a domain differs across different sections.

Introduction

The history of citation analysis dates back to 1955, when Eugene Garfield argued for the inclusion of citation indexing for scientific research (Garfield, 1955). Since then, citation indexing has found widespread adoption, facilitating the identification and ranking of both highly cited ideas, documents and authors, and by implication, their impact and importance. Aside from the bibliometrics that are enabled by counting citation events, there have been a variety of publications that go beyond simple counting. The core idea is to leverage the citation itself as an element with semantics, containing meaning that encodes the perspective of citing author, and providing a piecewise viewpoint of the intellectual structure of the domain.

In a related vein, document structure is a key element in scientific discourse, as scientific publications adhere to formalized conventions for disclosing research embodying the scientific method.

In scientific discourse, hypotheses, prior related work, methods and experimentation are presented in conventionalized argumentation style following formalized sections. Connecting this observation back to citation analyses, a research question that arises is whether the different patterns of highly-cited authors, when broken down by document sections, could be indicative of different roles in the development of a discipline. Do highly-cited authors in a field have their citations uniformly distributed over the different sections of the cited documents? Are the interpretable semantics as to how the author contributes invariant to the citation patterns of their works, when broken down by sections? To the best of our knowledge, there has not been any attempt to identify the characteristics of cited authors according to such logical document structure (i.e., *section,* or equivalently in this work, *location*), which would be able to answer such questions. We believe the answer to the above two research questions are "no", motivating work in the area of document location-sensitive citation analysis to measure their extent and effect. We believe that location-sensitive citation analysis enables novel insights into the characteristics of highly-cited authors, by refining our understanding of their contributions to a more fine-grained level of analyses. Our work aims to codify, formulate and explore this area of location-sensitive citation analysis by exploring 1) whether distinct characteristics do exist among location-specific citations patterns of highly-cited authors and 2) whether and how the citation location impacts the author citation network, which maps the knowledge structure of the given field, when built using per-author citation context similarity. To operationalize this research, we thus define *citation location* as the location of a citation as reported in the conventionalized sections of Introduction, Abstract, Background, Related Work, Method, Evaluation, Discussion, and Conclusion. By conducting a macro-level analysis of citation locations, we can identify major researchers cited on per section basis and understand their characteristics. In addition, by conducting detailed, per-author micro-level analysis of citation contexts, we can construct location-specific author networks. In detail, we apply advanced text mining techniques to calculate semantic similarity on our datasets and conduct both citation location analysis and citation content analysis.

To conduct such a study, it is necessary to confine our area of examination to a single domain within the scholarly literature, where expert assessment on the cited authors is obtainable and where the academic literature is openly accessible. To this end, we chose the discipline of computational linguistics and natural language processing, whose literature is largely represented by the single publisher of the Association of Computational Linguistics (ACL), and whose liberal copyright policy using the *Creative Commons 4.0 By* attribution license has allowed its scholarly literature to be freely accessible for study without licensing encumbrance. Furthermore, the *ACL Anthology Reference Corpus* (ACL ARC; Bird *et al*., 2008) and its recent large-coverage update converted much of the association's scholarly literature into a full-text scholarly corpus, which we utilize for our study. The updated ACL ARC represents over 30,000 scholarly conference and journal articles, of which the majority are conference papers. We note that in the field of computational linguistics, conference papers are the primary form of research communication, whereas journal publications are used both for archival and as extended versions of the primary research.

The major findings of our work are three-fold: First, a different set of highly cited authors share distinct academic characteristics according to their citation locations. Second, the citation locations of the majority of highly-cited authors differ by ranking percentile. Third, there are differences among the author network structure when constructed using different sections as the bases for construction. In addition, by conducting detailed, per-author micro-level analysis of citation contexts, we construct location-specific author networks. In detail, we apply advanced text mining techniques such as model-based XML parsing and window size optimization to calculate semantic similarity among citation contexts and conduct both citation location and citation content analysis.

Related work

Citation location analysis

We use the term "*location*" in this work interchangeably to connote logical document *sections* that commonly divide an academic paper into its constituent components, and "*position*" to denote the word offset within a scientific work. Voos & Dagaev (1976) were the pioneers who first suggested that citation locations have different implications. They insisted that the location of a citation, in terms of the logical document structure – such as in the Introduction, Methodology, Discussion and Conclusion – has to be considered as part of the citation element for evaluation, rather than only applying simple counting to arrive at citation frequency. In the same vein, Herlach (1978) assigned higher weight to citations appearing in several positions within a paper as compared to citations that only appear in one position. McCain & Turner (1989) assigned different values to citations with respect to citation location (specifically, Introduction, Method, and Discussion).

In recent research, these previously manually-assessed techniques have been automated. Gipp & Beel (2009) revisited citation position evidence, arguing that relevance of authors who are co-cited are influenced by whether they are cited in same position or not. With this assumption, they invented the *Citation Proximity Index* that assigns a larger value to co-cited work cited at same position. They conducted a pilot task on relevant document search using the co-citation relation that validated that the use of the Citation Proximity Index outperforms the use of simple co-citation index, when only the co-citation relation is used for document indexing. However, the impact of the locations of the citations was not dealt with in their study as they calculated proximity of cited documents using only positions (word offsets). In a similar vein, Elkiss et al. (2008) devised a similarity metric to build a citation summary representing a cited paper. They considered sections as document units to calculate the similarity between co-cited papers. However, their work did not focus on sections and merely used it as one of the criteria to measure structural proximity. Ding *et al.* (2013) successfully verified the pioneering idea of Voos & Dagaev (1976) by incorporating text mining techniques into citation analysis. They identified the location distribution of highly-cited documents and compared the rank as determined by simple count with the location-sensitive rank, generated by considering the occurrence in different locations of a document. Zhu *et al.* (2015) intuitively identified the different important aspects of citations from different sections.. They used citation location as one of the features in their model to measure the impact of a cited work. Hu *et al. (*2013) studied the distribution of citations within the document structure. They found that even when citations are uniformly distributed within a document, different sections have unequal probability of hosting citations. A key limitation of their study is that their document structure boundaries are determined from the length of the document. They divided a document into four parts according to its length and assigned each the four parts the prominent four document sections, namely, Introduction, Methods, Results, and Conclusions. However, they also disclose that many documents in their corpus consisted of more than four logical parts, motivating a need for a more precise method to detect sections of a document. Moreover, all of the aforementioned works deal with small sized datasets, as their analyses were done manually. Although this is understandable, invalid to extrapolate their findings to scale and statistical significance. Therefore, in this study, we use automated (and occasionally error-prone) methods to detect the structure of a document automatically, and apply our analyses to a sufficiently large dataset. This allows us to justify this study models the reality of certain, actual scientific publications within the limited domain of computational linguistics.

Citation content analysis

Citation content analysis is the one form of citation analysis using “the semantic content of the citing passage to characterize the citing work” (McCain & Turner, 1989). The citing passage is termed the “citation context”, and according to Small (1982), citation *content* analysis is disentangled from citation *context* analysis, in the differentiation of its analysis goal. Citation *context* analysis has a broader meaning than *content* analysis in that it attempts to understand the perspectives of the citing author that is embodied by the content of the context. Lipetz (1965) conducted seminal work in establishing a classification scheme by analyzing citation contexts. He analyzed 750 citations from 60 papers, suggesting a typology of 29 citation relationships. Subsequently, there were several attempts to code a citing author’s perspectives appearing in citation context (Moravcsik & Murugesan, 1975; Moravcsik *et al.*, 1976; Murugesan & Moravcsik, 1978; Shearer & Moravscik, 1979). These studies influenced automatic classification of citation context (Garzone & Mercer, 2000; Teufel *et al.*, 2006). In another line of research, citation content analysis focuses on the subtle conceptual meaning of cited work, as derived from Garfield’s (1970) earlier work. He applied the concept and theory proposed by Merton, a celebrated scholar in social sciences, to the scholarly communication domain. Garfield regarded the citation contexts as means of describing cited works such as subject terms. Small (1978) applied this notion concretely to find core concepts in the domain of chemistry by analyzing content of citation contexts. He thought that the citation context is the symbolic expression of the ideas in the cited work made, as expressed by the citing author. This interpretation of citation context is used in McCain & Turner (1989), who conducted content analysis on citations to identify the change of important concepts from key papers in molecular genetics over time. McCain & Salvucci (2006) used citation content analysis to observe how a certain theme in the field of software engineering was treated in other fields. Siddharthan & Teufel (2007) treated citation functions as attributes of the cited works and applied machine learning techniques to automatically identify such citation functions. Small (2011) elaborated the citation content analysis to draw an intellectual map of science. He conducted sentiment analysis on citation context using linguistic tools and plotted the geography of the science domain as co-citation network. Recently, research comparing the efficacy of co-citation analysis reflecting with and without contents has also been conducted (Jeong *et al.,* 2014). Their study revealed that the co-citation network that considers citation content yields a higher resolution intellectual map in the domain of information sciences. A common finding of these studies is that distinguishing citation sentences from citation context is ambiguous or arbitrary. Sometimes, two or three sentences around the citation expression are considered as context of citation. In addition, selection of context is manually done. In contrast, in this paper, the boundary of context is detected automatically for precise citation content analysis and to enable analyses at scale, as described in the next section.

Method

Our author network construction pipeline consists of three phases; Data Collection, Pre-processing and Data Analysis (Figure 1). First, we collect and pre-process the raw data to make it suitable for downstream analyses; and then compute basic descriptive statistics, and the semantic similarities of authors' cited contexts. The similarities are finally used to build the author networks for each section of a document.

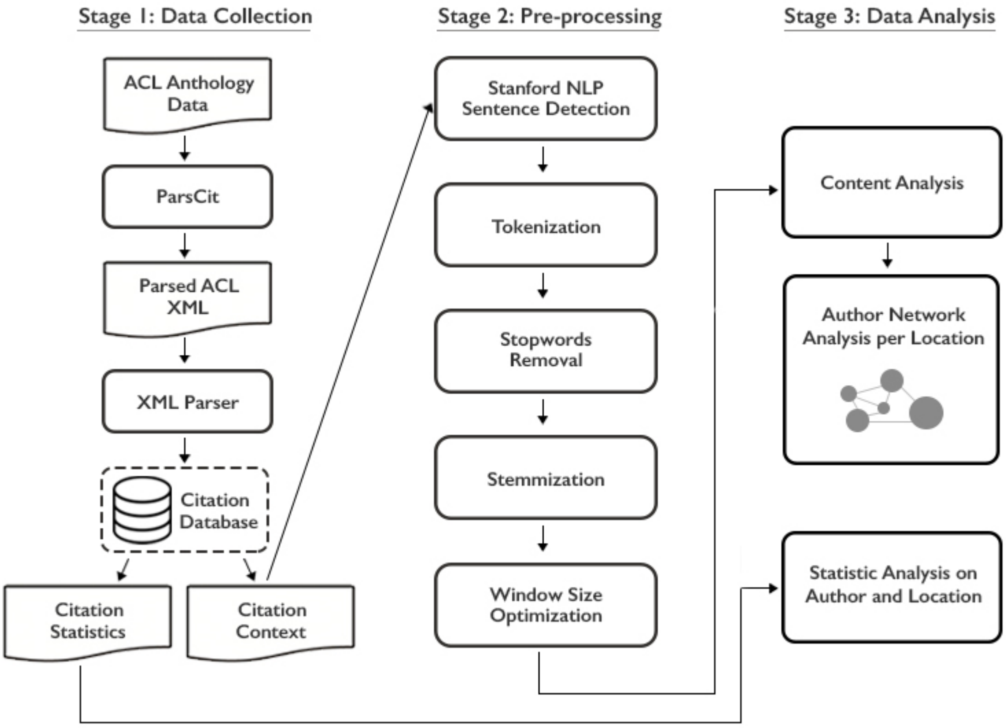
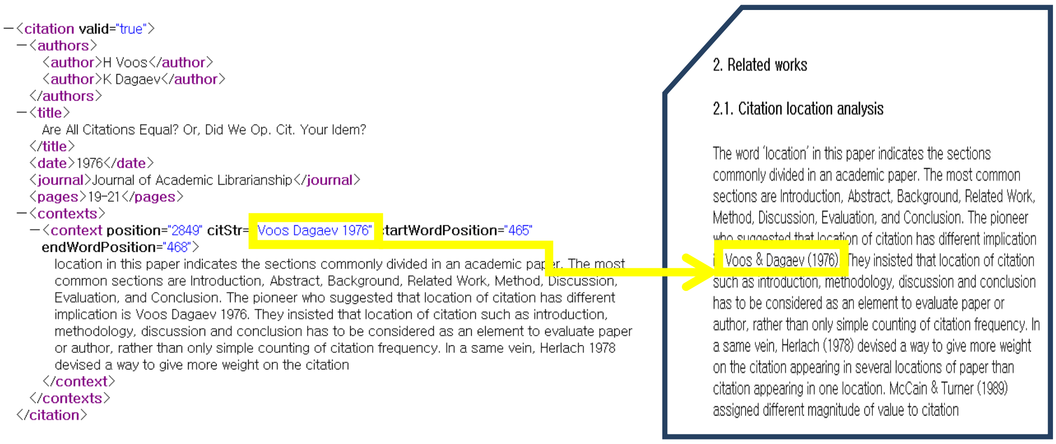


FIG. 1. Overview of our 3 stage pipeline to construct author networks

Data collection

As described in brief earlier, the dataset used in our study is derived from the ACL Anthology Reference Corpus (Bird *et al*., 2008), which hosts over 34,000 papers on the study of computational linguistics and natural language processing (hereafter, CL/NLP). From the recent version 2 update of the ACL ARC, we extracted 29,158 full-text papers in eXtensible Markup Language (XML) derived using the commercial Nuance Omnipage (Versions 16 and 17), an OCR software package, and fed them to ParsCit (Councill, 2008; Luong *et al*. 2010), an open-source package that parses the logical document structure and bibliographic reference strings with a series of trained conditional random field (CRF) models. For the purpose of our work, the aspect of ParsCit that is most critical is that it identifies and provides a unified citation context per citation in a document, output in XML. This makes it possible to maintain a consistent level of quality in our citation context analyses at scale. Figure 2 shows an example of a parsed XML using ParsCit.



```
—<citation valid="true">
  —<authors>
      <author>H Voos</author>
      <author>K Dagaev</author>
    </authors>
  —<title>
      Are All Citations Equal? Or, Did We Op. Cit. Your Idem?
    </title>
    <date>1976</date>
    <journal>Journal of Academic Librarianship</journal>
    <pages>19-21</pages>
  —<contexts>
    —<context position="2849" citStr="Voos Dagaev 1976" startWordPosition="465"
      endWordPosition="468">
        location in this paper indicates the sections commonly divided in an academic paper. The most
        common sections are Introduction, Abstract, Background, Related Work, Method, Discussion,
        Evaluation, and Conclusion. The pioneer who suggested that location of citation has different
        implication is Voos Dagaev 1976. They insisted that location of citation such as introduction,
        methodology, discussion and conclusion has to be considered as an element to evaluate paper
        or author, rather than only simple counting of citation frequency. In a same vein, Herlach 1978
        devised a way to give more weight on the citation
      </context>
    </contexts>
  </citation>
```

2. Related works

2.1. Citation location analysis

The word 'location' in this paper indicates the sections commonly divided in an academic paper. The most common sections are Introduction, Abstract, Background, Related Work, Method, Discussion, Evaluation, and Conclusion. The pioneer who suggested that location of citation has different implication is Voos & Dagaev (1976). They insisted that location of citation such as introduction, methodology, discussion and conclusion has to be considered as an element to evaluate paper or author, rather than only simple counting of citation frequency. In a same vein, Herlach (1978) devised a way to give more weight on the citation appearing in several locations of paper than citation appearing in one location. McCain & Turner (1989) assigned different magnitude of value to citation

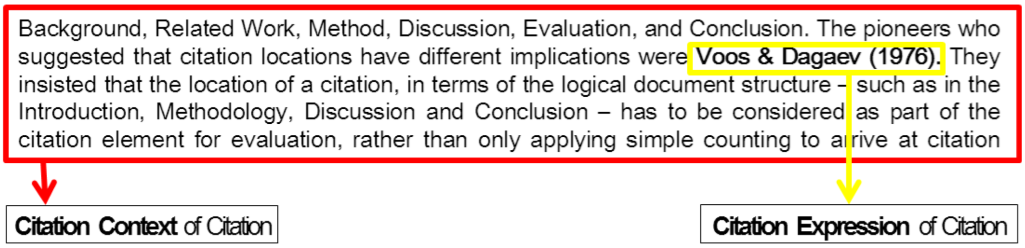
FIG. 2. Example of the output XML obtained from ParsCit (left), and its input raw document (right).

The full-text articles in XML format were further processed by a SAX-based XML parser we developed. Raw XML and processed data are then stored in our citation database to facilitate random access and downstream analyses. Logical document structure of documents is obtained using SectLabel (Luong *et al.* 2010) module in the ParsCit suite. SectLabel extracts and labels logical structures in a paper such as author and affiliation lines. SectLabel maps the section's named title to one of the 13 generic section types that SectLabel supports (e.g., "Document Preprocessing" would be mapped to "Methods"), which is important in handling idiosyncratic section names. Although SectLabel is able to detect 13 types of logical document sections, we limit the types of sections we consider to eight – namely, Introduction, Abstract, Background, Related Work, Method, Evaluation, Discussion, and Conclusion – as these section types constitute the majority of the sections detected in the ACL ARC. Documents are parsed with the section information according to the eight types. The parsed content is saved in our database according to the XML tags on original XML files. We built a citation database from the source 29,158 articles, resulting in 466,366 citation contexts, or roughly 16 in-body citation contexts per scholarly paper. We believe this number is in the normal range for citation contexts for conference paper works, where the body of the paper has strict page limits (usually 8 pages), and validates the reasonableness of the extraction process. The total number of raw (i.e., without author disambiguation and normalization) authors in the entire dataset is 101,041, or 3.46 authors per paper on average.

Pre-processing

We used Stanford CoreNLP (Manning *et al.* 2014), an open-source Java-based natural language processing toolkit, to perform much of the preprocessing steps on the parsed text. After determining the language of the article to be English using language detection techniques (Nakatani, 2010), we sentence-delimit and tokenize sentences using CoreNLP, and subsequently remove stop words and special characters. We finally use the Porter stemmer algorithm to conflate common term variations (Porter, 1980), to calculate context similarities.

Determining the citation context boundaries



Background, Related Work, Method, Discussion, Evaluation, and Conclusion. The pioneers who suggested that citation locations have different implications were **Voos & Dagaev (1976).** They insisted that the location of a citation, in terms of the logical document structure – such as in the Introduction, Methodology, Discussion and Conclusion – has to be considered as part of the citation element for evaluation, rather than only applying simple counting to arrive at citation

Citation Context of Citation
Citation Expression of Citation

FIG. 3. Citation expression and Citation context of Citation

In this study, a *citation* refers to a position in a citing document where the citing author references the cited paper. In a citation, a *citation expression* is used to make the reference – typically using surnames, year of publication or an index number, delimited with some form of bracketing punctuation. The citation expression cross-references a bibliographic entry for the cited paper at the end of the document, in a bibliography or reference section. All of the citation expressions within a document typically follow a publication venue regimented style. This regularity makes citation identification programmatically easy. Deciding the context of a citation – i.e., the portion of the text that is relevant to the citation – is more difficult, as the context can vary per citation and can be semantically ambiguous. Figure 3 describes an example of a citation context, illustrating how sentences both before and after the citation expression can serve as context. As discussed in previous works, studies on citation context have been done manually and determined the boundaries of citation contexts laboriously, which does not admit the feasibility of large scale analyses. In contrast, we determine the citation context boundary using NLP based methods over potential citation contexts and choosing a setting that works optimally. In a previous work to determine the citation context boundary, Qazvinian & Radev (2010) applied graphical models to citations extracted from 10 scientific papers and compared their output to the gold standard. They found the performance to be optimal when using four surrounding sentences before and after citation sentence (two before, two after). Meanwhile, they adapted cosine similarity under the intuition that sentences which have high similarity with the cited paper have a high probability of being part of the correct citation context. The setting that gives the optimal task performance is when one sentence both before and after the citation (a total of two) are used as the citation context. Also, their experiments were conducted at a document level rather than at section level.

Starting with a citation expression, we segmented the nearby text into sentences as determined using the CoreNLP sentence delimiter. We then calculated the average similarity value for each type of logical document section using the JaroWinklerTFIDF algorithm to decide an optimized window size for the citation context, where the window is centered on the sentence hosting the citation expression. For example, a window size of 1 means that there are 2 sentences around the sentence hosting citation expression. The window sizes we experimented with are 1 (smallest), 2, and 3 (largest). We hypothesize that if the text within a certain window size of citation context shows high intra-text similarity, it indicates than the window's size is a better fit of the actual citation context, per section.

The JaroWinklerTFIDF algorithm combines the JaroWinkler algorithm with the cosine similarity algorithm (Cohen *et al.*, 2003). The JaroWinkler distance algorithm was originally designed to match variant name and address records generated from US census data. The algorithm makes it possible to assign partial weight to a word pair according to the number of common characters (Jaro, 1989), as shown in Equation 1 below:

$$d = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m}\right) & \text{if } x < 0 \end{cases} \qquad (1)$$

$$d_{jaro-winkler} = d_{jaro} + \ell P(1 - d_{jaro}) \qquad (2)$$

where m is the number of common characters in two tokens, S_1 and S_2 indicate the number of characters of tokens 1 and 2, respectively and t is the average number of common character(s) which have different indices in the two tokens. These character differences are summed as d and fed to Equation 2, which weights this difference against the prefix similarity. An example of JaroWinkler similarity computation is shown in Figure 5 – here the final similarity is higher than the initial edit similarity calculation ($d = 0.944$), as the common prefix is long (final $d = 0.967$).

Token 1	A	P	P	L	I	C
Token 2	A	P	P	L	I	

$$\left.\begin{aligned} m &= 5 \\ |S_1| &= 6 \\ |S_2| &= 5 \\ t &= 0 \end{aligned}\right\} \; d = 0.944,\;\; d_{jaro-winkler} = 0.967 \text{ (p = 0.1)}$$

FIG. 4. Example of JaroWinkler distance calculation

After calculating distance between tokens, if the intra-text similarity is high (value larger than 0.7), the text is used to construct a term vector to calculate cosine similarity between sentences according to Equation 3. Figure 5 summarizes our similarity processing in a nutshell, illustrating the process of calculating cosine similarity and compares the similarity obtained by using the stemmed, stopworded JaroWinkler preprocessing against ordinary simple count for cosine similarity: one can see that our JaroWinkler cosine similarity better reflects the high degree of semantic homogeneity between the two sentences.

$$\mathrm{S(A, B)} = \frac{\sum_{i=1}^{t} a_i b_i}{\sqrt{\sum_{i=1}^{t} (a_i)^2 \cdot \sum_{i=1}^{t} (b_i)^2}} \qquad (3)$$

Sentence A: This method is applied to preprocess textual data
Sentence B: Application of this method is used for text pre-processing

- *Simple Frequency Term Vector* : Cosine Similarity = 0.472

	Applic	method	appli	preprocess	text	textual	data
Sen A	0	1	1	1	0	1	1
Sen B	1	1	0	1	1	0	0

- *Jaro Winkler Term Vector* : Cosine Similarity = 0.868

	method	appli, Applic	preprocess	text, textual	data
Sen A	1	0.967	1	0.914	1
Sen B	1	0.967	1	0.914	0

FIG. 5. Example of cosine similarity calculation

We applied this technique to optimize window sizes per document location over all 466K citation contexts in the ACL ARC dataset. Table 1 shows the intra-citation context similarities for contexts with different window sizes between 1-3; *i.e.*, with a total 3 to 7 sentences in the citation context, as well as a default setting where all of the sentences within the citing context extracted by ParsCit automatically according to a probability score are used as the context.

TABLE 1. Per-location intra-text context similarity for different context sizes

	All sentences around the citation sentence	(Window size 1) 2 sentences around the citation sentence	(Window size 2) 4 sentences around the citation sentence	(Window size 3) 6 sentences around the citation sentence
Introduction	0.143	**0.148**	0.146	0.141
Abstract	0.134	**0.151**	0.151	0.142
Background	0.133	**0.134**	0.133	0.133
Related work	0.151	**0.153**	0.153	0.148
Method	0.132	**0.140**	0.137	0.131
Evaluation	**0.130**	0.127	0.128	0.125
Discussion	0.132	**0.133**	0.133	0.127
Conclusion	0.130	**0.135**	0.132	0.126

For most of locations, the highest similarity value (bolded) is achieved when the window size of the context is 1. There is an exception, where in the case of Evaluation sections, the highest value occurs when all sentences in citation context are considered as boundary of citation context, but the difference between the optimal value and a window size of 2 is small. As such, we have uniformly applied a window size of 1 (*i.e.*, the sentence hosting the citation expression, as well as the sentences directly before and after, when applicable) for all document sections in the remainder of this work.

Content analysis using citation context

To explore the characteristics of highly-cited authors in each section at a macro level, we calculated similarity among citation contexts of the top 100 cited authors by document section. We take the simplifying assumption that all authors of a paper contribute equally towards the paper in our work. We then applied the aforementioned JaroWinkler distance algorithm to calculate similarity of selected authors' citation contexts as illustrated on the left side of Figure 6. The calculated similarity between citation contexts belonging to different authors are used (after summing) as weights for edges between the said authors in building an author network. We built an author network for each of the eight sections to compare distinct characteristics of author network according to the cited location in a document.

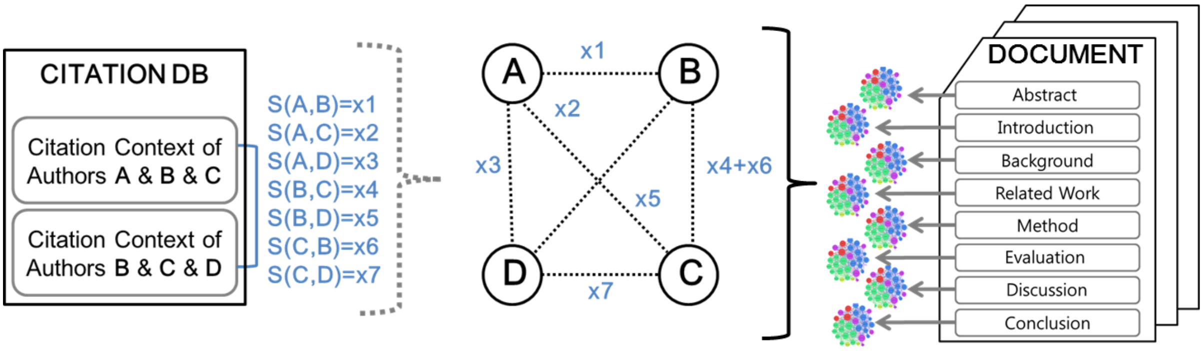


FIG. 6. Workflow to construct author networks

The resultant networks' nodes are authors and edges reflect the similarity of citation contexts. Edge strengths are weighted by the similarities between citation contexts of the authors. To visualize the author networks, we employed Gephi 0.8.2. (Bastian *et al.*, 2009), setting the size of (author) nodes relative to their degree centrality, and coloring the nodes to show the authors' induced community as computed using the modularity algorithm (Blondel *et al.*, 2008), with resolution set to 0.9. In other words, if the two nodes have the same color in the visualization, they have high similarity in their citation contexts. To expedite the analysis of each community, we also annotated the sub-discipline of the community based on the authors being included in each community.

Result

TABLE 2. Simple statistics of citation distribution in the ACL ARC (Numbers in parenthesis are percentages)

Section	Number of Citation	Number of Author-Citation pair
Introduction	116,966 (37.29)	720,090 (36.97)
Abstract	8,842 (2.82)	42,808 (2.20)
Background	1,080 (0.34)	6,334 (0.33)
Related work	28,552 (9.10)	173,413 (8.90)
Method	135,125 (43.08)	847,831 (43.52)
Evaluation	15,788 (5.03)	117,748 (6.04)
Discussion	1,040 (0.33)	5,456 (0.28)
Conclusion	6,300 (2.01)	34,245 (1.76)
Total	313,693 (100.00)	1,947,925 (100.00)

We first give an overview of the characteristics of the citation distribution of cited authors with relative to their sectional location. We examine the top authors by their sectional rankings and examine how these per-author distributions vary as the author citation ranking decreases. We then discuss rankings per section, highlighting our analyses using specific instances to support our arguments.

As a starting point, the demographics of citation contexts (Table 2) yields useful information about the dataset. Recall in Method that the dataset consists of 29K articles, with 466K citation contexts that have been extracted automatically. After attributing each context to possibly multiple authors, we recover 1.9M (1,947,925 to be exact) recognized author-citation context pairs in the dataset. This yields 66.8 author-citation context pairs per cited paper and 6.2 authors per cited context, on average. We note that the average number of authors varies slightly from the previous figure of 3.46 earlier (as derived directly from the totals number of raw authors and papers), due to the dataset used here is restricted to citation from eight sections and normalization process we employed. It is impossible to deal with all names of authors and that is not our purpose of the study, so we just disambiguate the top-cited authors' name manually, building the list of 100 top-cited authors per section.

The figures in Table 2 validate prior work (Cano, 1989; Voos, & Dagaev, 1976) that the majority of citations appear in the front portion of scholarly works. We note the distinct peak in citation contexts that are attributed to (possibly multiple) Method sections in the papers. As stated earlier, the ACL ARC dataset primarily consists of conference articles that describe recent innovations in CL/NLP, and the bulk of the prose body in these papers are attributed to Methods section. It is thus natural that the largest share of citation contexts occurs in the most prevalent section.

Top 100 authors of each location

We now proceed to the central findings of our study, the analysis of highly-cited authors in the ACL ARC network. Using the 1.9M author-citation context pair tuples, we calculated per-section rankings of authors, over the eight prominent section types discussed earlier. As to be expected, the distribution of highly-cited authors is not uniform, and it could be assumed that the distribution follows a power law distribution because half of the authors (50,806) are cited less than twice in the data set. Without loss of generality, we truncate the rankings to the top 100 most highly-cited authors per section, and the subsequent analyses focuses on these authors and their associated citation contexts.

279 distinct authors appear in the top 100 ranks in the rankings for the eight locations. 11 authors are highly cited (in the top 100 most cited) over all eight sections (visualized per location in Figure 7). Within the top 50 ranks, three authors (Chris Manning, Fernando Pereira and Philipp Koehn) of the initial 11, remain highly cited in all sections. Additionally, four authors (Chris Callison-Burch, Eugene Charniak, Mark Johnson and Philip Resnik) are ranked in the top 50 with the exception of just one location. The 11 authors represent some of community's most important authors. For example, Chris Manning, at rank 1, is the author of one of the standard textbooks on computational linguistics and was the president of the association in 2015. Separately, 95 (34.0%) of the authors in the top 100 ranks are cited in only one of the eight sections.

Tabke 3 lists top 50 authors that are cited in only a single section. Note that in this table, there are no authors cited exclusively in the Method section. That is, all of the Method sections' 50 authors are referred in other sections at least one time. In other words, in terms of both size and top authors, the Method section is the most representative section with citations in the CL/NLP domain. With the locations that occur less frequently (i.e., Background and Discussion), we see many more names than in the common section types. The names represent, on average, more junior contributors to the field, which may be present possibly due to the sparsity and noise in the data. Meanwhile, for the Evaluation and Conclusion section, there are just two authors, even though the proportion of citation context relative to the whole set of citation contexts is just 5% and 2%, respectively. It shows that, even if it is restricted to just top 50 authors, the works cited in Evaluation and Conclusion section are done by general top authors and are likely to contain classic content, not newer advances.

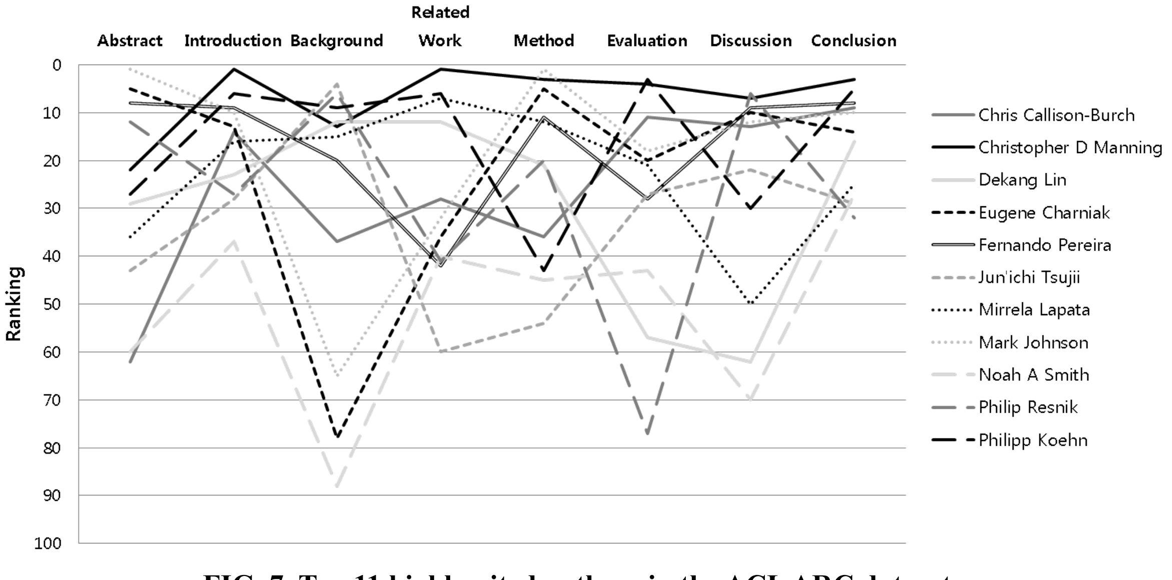


FIG. 7. Top 11 highly-cited authors in the ACL ARC dataset

TABLE 3. 41 of the Authors cited in the top 50 ranks are cited in only one section

Location	Authors (number)
Abstract	Stephan A. Della Pietra, Stuart M. Shieber, William A. Gale (3)
Introduction	Fredrick Jelinek (1)
Background	Brian Murphy, Ching-Yun Chang, Christian Grothoff, Els Lefever, Idan Szpektor, Johan Hall, Mark Lauer, Mikhail J. Atallah, Veronique Hoste, Yoshua Bengio, Rodger Kibble, Michael White, Katrin Erk, Zheng Chen (14)
Related Work	Vasileios Hatzivassiloglou, Fabrizio Sebastiani, Ming Zhou (3)
Evaluation	Wade Shen, Christine Moran (2)
Discussion	Marilyn A. Walker, Anna Korhonen, Mark Dras, Eric Joanis, Shervin Malmasi, Edward Gibson, Sabine S. Walde, Jphanna D. Moore, Herbert H. Clark, Roxana Girju, Joshua B. Tenenbaum, Jean Carletta, Rasimi Prasad, Julia Hirschberg, Nathanael Chambers, Yuval Krymolowski (16)
Conclusion	Julia Hockenmaier, Julie Weeds (2)

Citation trends by location with respect to author ranking

Hypothesizing that top-ranked authors have distinctive citation-location patterns, we examine citation-location distributions for successively relaxed partitions of top *n* most-cited authors (i.e., in aggregate, over all locations). At each rank, authors are designated to one section where they most frequently appeared. Figure 8 shows a sequence of citation-location distribution charts, for the top 10, 20, 30, 50, 70 and 100 ranked overall cited authors.

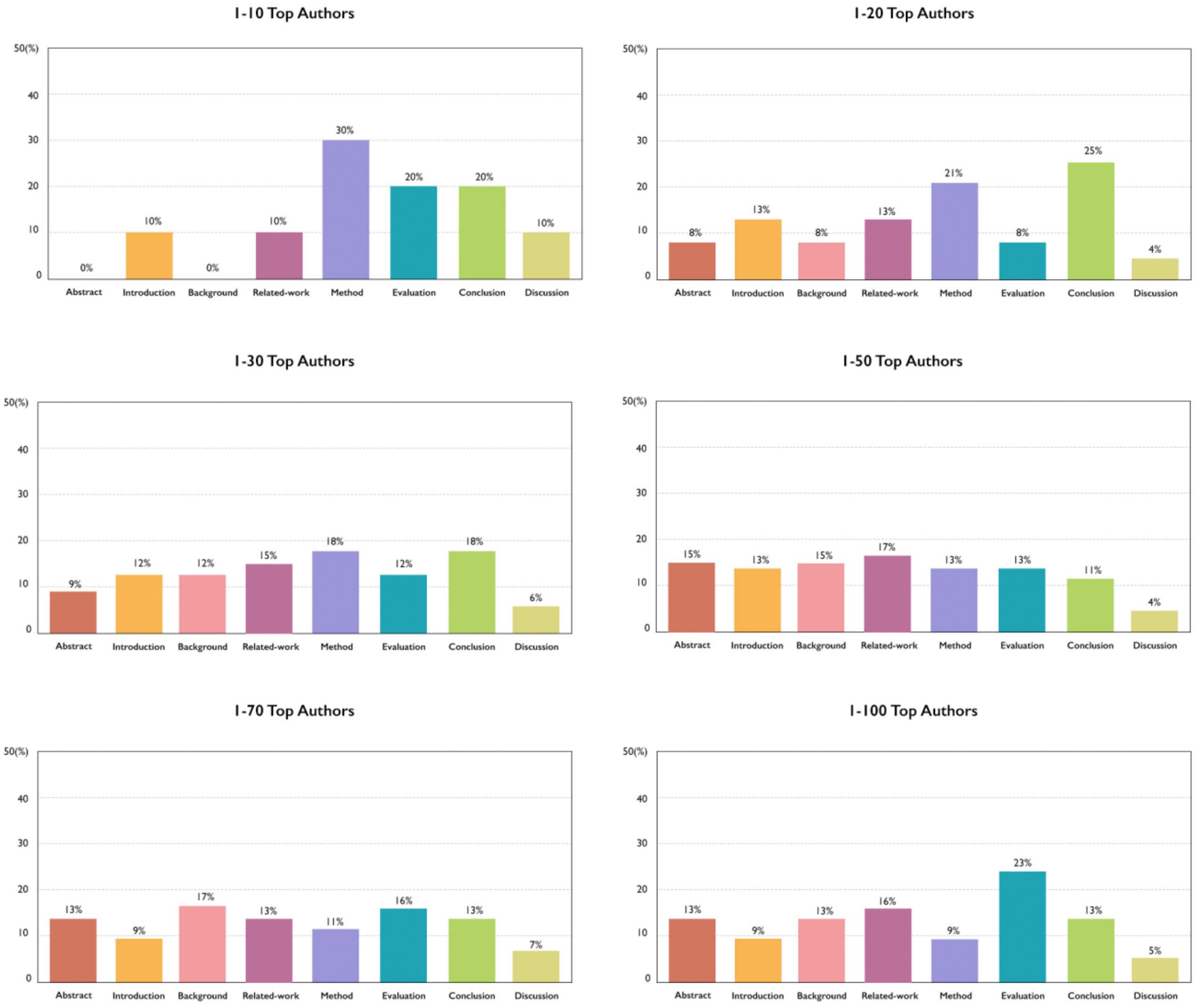
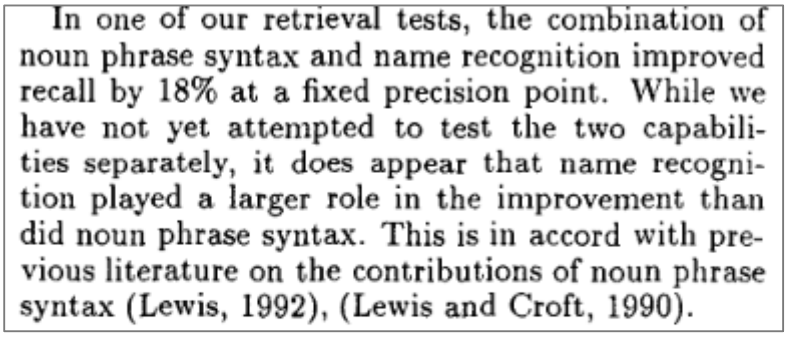


FIG. 8. Top 10 to top 100 authors' citation-location ratios

An interesting pattern emerges from the sequence of charts. Top ranked authors have a clear peak in citations in the Method location of citing papers. The CL/NLP community cites these authors in the Method section possibly for a variety of reasons: to connect their work as using established methodologies communicated by these authors or to compare their methodologies with these existing ones. This pattern dampens a bit as the criterion for citation inclusion is relaxed to encompass other authors; at the top 30 ranks, the Conclusion section also figures prominently in the citation-location distribution, perhaps illustrating the citation ratio in Conclusion. Citation within the Method section remains significant as well. Similar to the case of the Method section, important authors are cited frequently in the Conclusion section, perhaps to align their findings to popular and known projects in the field.

Another interesting observation comes from considering trends apparent over all of the graphs' patterns. The proportion of citations within Evaluation sections increases across all graphs, with the exception of the first (1-10 top authors) graph. Accumulating all graphs (*i.e.*, top 10-100), Evaluation becomes the section with the largest share of citations. Evaluation is a part where scholars estimate the result of their studies and support their own assertions. Therefore, we may hypothesize that they cite well-recognized works (and by association, authors) to strengthen their research.

To sum up, the top highly-cited authors are mainly cited in the Method and Conclusion sections, perhaps to be compared or consulted. However, as we broaden the scope to encompass all top 100 authors, the Evaluation section shows the highest ratio, possibly to support the result and validity of research across many specific sub-domains. For example, in Figure 9, the work of authors who have contributed to the domain of citing paper is referred in the Evaluation section.



In one of our retrieval tests, the combination of noun phrase syntax and name recognition improved recall by 18% at a fixed precision point. While we have not yet attempted to test the two capabilities separately, it does appear that name recognition played a larger role in the improvement than did noun phrase syntax. This is in accord with previous literature on the contributions of noun phrase syntax (Lewis, 1992), (Lewis and Croft, 1990).

FIG. 9. An example of (Flank, 1998, p.402) citing in its Evaluation section

Author networks by content analysis of each location

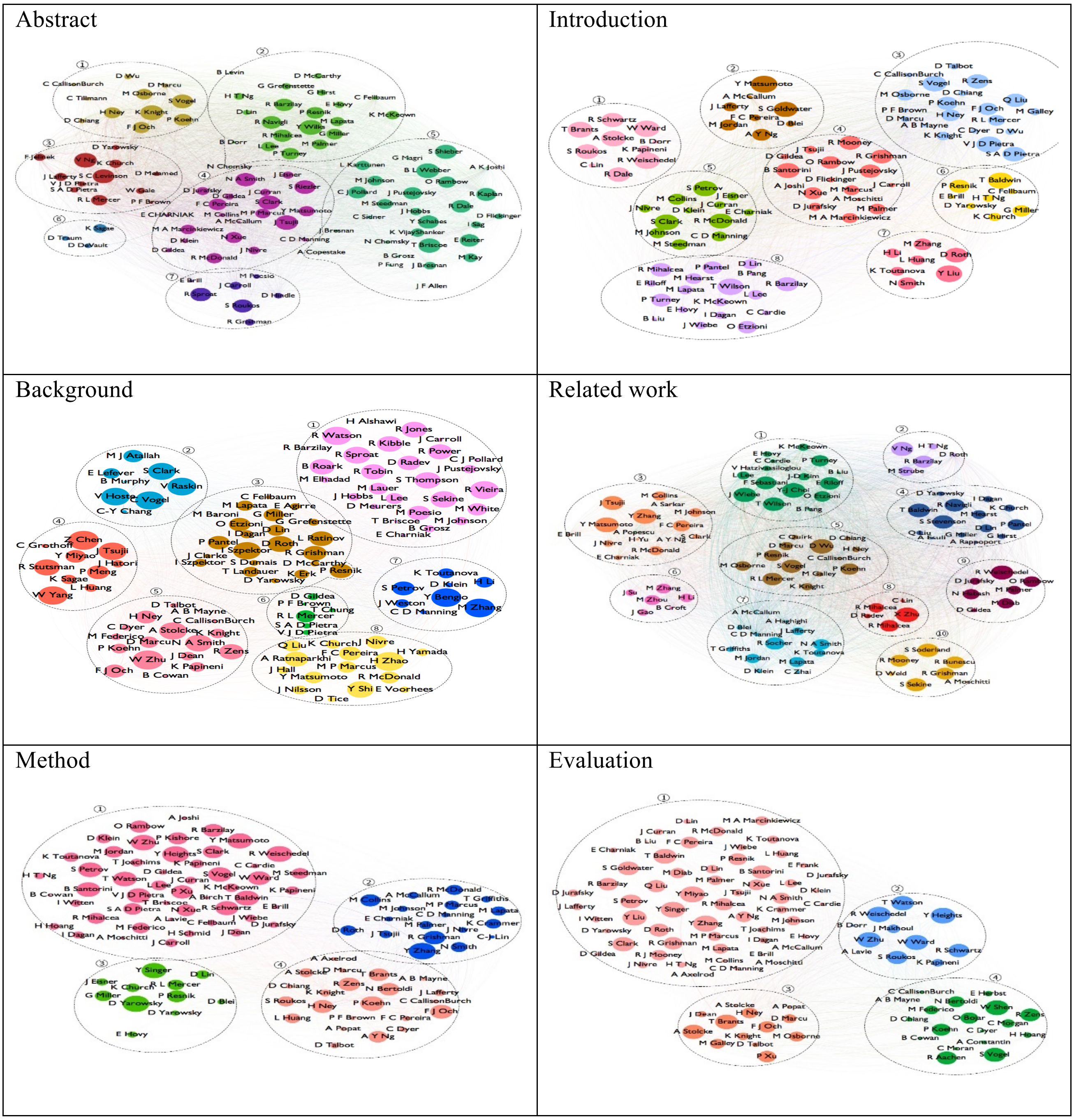
As described in Figure 5, we built author networks for each section using semantic similarity among citation contexts of authors. In related work, Joseph & Radev (2007) and Radev *et al.* (2009) analyzed top-ranked journals, papers and authors using network metrics such as clustering coefficient and PageRank. These statistics can be readily computed from a corpus as demonstrated by the ACL Author Network (AAN)[1]. The analysis of the AAN produced metrics that were then compared to Impact Factor. Weitz & Schäfer (2012) modelled papers as nodes and citation relation as edges to render a graphical citation browser of the ACL Anthology. However, none of the prior works provide a visualization of the author network that reveals the intellectual sub-structure of the domain. Furthermore, our network visualization clusters authors who are cited in similar vein. Figure 10 shows eight of the section-specific author networks[2]. Table 4 shows the results from section-wise network analysis. We map the communities to detect various sub-disciplines of CL/NLP based on the works by the authors in each community. Sections with sparse citation contexts (Abstract, Background, Discussion, Conclusion) are not labelled. Citation contexts from Method and Evaluation sections fall into broader disciplines that branch into multiple sub-disciplines than citation contexts from Introduction and Related work sections. This could be attributed to the general nature of citations in the Method and Evaluation sections that map to broad sub-disciplines.

The semantic similarity among citation contexts represented by the edge weights of author networks is an average over all the nodes in the network. The average weighted degree for each node (author) is normalized by the number of citations attributed to the citation-author pair. It is found to be the highest in the Evaluation section. We conclude that the works cited in the Evaluation section are referred in more similar contexts than in other sections. This holds for the Background section too.

[1] http://clair.eecs.umich.edu/aan/index.php

[2] More detailed figures and analyses are available upon publication.

(Normalized Average Degree: 42)



Abstract

Introduction

Background

Related work

Method

Evaluation

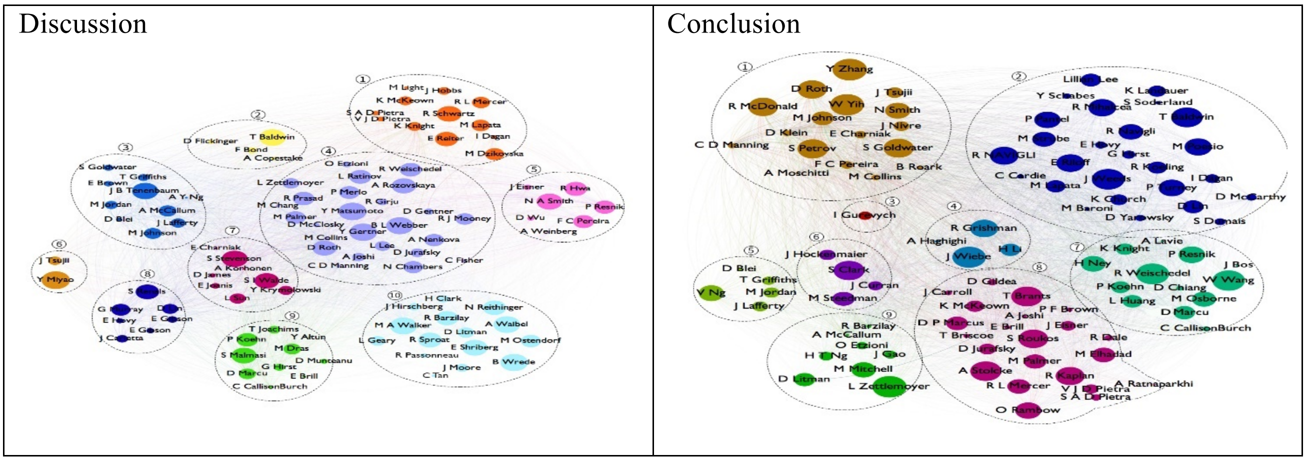


FIG. 10. Author networks

TABLE 4. Sample description of author networks

	Introduction	Related work	Method	Evaluation
Number of Communities detected	8	10	4	4
Sub-disciplines	1 Machine Learning 2 Speech Processing 3 Machine Translation 4 Lexical Resources 5 Core NLP 6 Lexical semantics 7 Statistical Models 8 Information Retrieval, Extraction and Sentiment Analysis	1 NLP Applications 2 Statistical NLP 1 3 Statistical NLP 2 4 Lexical Semantics 5 Machine Translation 6 China/Singapore-Affiliated NLP 7 Statistical NLP 3 8 Information Extraction 1 9 Human Languages and Resources 10 Information Extraction 2	1 General NLP 2 Core NLP 3 Statistical Models 4 Machine Translation	1 General NLP 2 Industrial NLP 3 Speech and Discourse 4 Machine Translation
Normalized Average Degree	32	32	25	65

Discussion

We juxtaposed the per-section citation distribution for all authors and those limited to the top 100 highly cited authors. Table 5 shows the number of top-100 cited authors (per section pair) who were cited in both sections. Recall that the Method section has the highest citation frequency, followed by Introduction (43.08% and 37.29% respectively). There are 74 authors in common between the two sections (see Table 5). This implies that overall citations occur a lot and that such highly-cited important authors are cited in Method and Introduction. Recall from Figure 8 that the top 100 authors are most cited in the Evaluation section, which has a relatively low proportion (5%; see Table 2) of the total citations. Even though the number of overall citations is small, the highly-cited authors are disproportionately cited in the Evaluation section. The co-occurrence statistics in Table 5 shows that almost 80% of authors ranked in top 100 for Evaluation section are also ranked for the Method section, giving further evidence to our claim that scholars frequently cite prominent authors in the Evaluation section to support their own assertion and strengthen the validity of their research.

TABLE 5. Number of top 100 authors of each section who co-occur in two sections (The two bolded figures indicate a high degree of co-occurrence and are discussed in the text.)

	Introduction	Back -ground	Related Work	Method	Evaluation	Discussion	Conclusion
Abstract	59	43	47	52	45	37	56
Introduction		46	66	**74**	71	40	67
Background			35	48	44	24	44
Related Work			-	53	51	39	61
Method				-	**78**	38	63
Evaluation						33	53
Discussion						-	40

To sum up, we infer that an author is cited in a certain location due to certain unique traits of their work. Typically, citations in the Introduction section serve to motivate the proposed study and survey existing related work. We may therefore infer that authors cited in Introduction act as 'motivators'. Citations in the Background section serve to inform readers about the background of the research, the major milestones and the state-of-the-art. Therefore, each citation in Background would be deeply subdomain-specific. This is validated by the relatively large number of communities in the author network, 8 (see Figure 10). Consequently, relatively 'peripheral' authors are ranked highly in Background section (Table 3). Citations in Related Work section tend to introduce other important researchers and their works who have influenced and shaped the domain, here the CL domain. This possibly explains why the Related Work section has just 3 exclusive authors (Table 3) and 10 communities (Table 4). Authors cited in Related Work may be inferred as "key players in the community".

Citations in the Method section acknowledge existing work or compare with existing methodologies. Recall that, no author is exclusively cited in the Method section (Table 3). This may imply that an author who is frequently cited for methodology is an eminent scholar and is likely cited in other sections too. Citations in Evaluation are made to assess and support the results of researches. In other words, the Evaluation section extends from the Method section. This is reinforced by the high co-occurrence of cited authors between the two sections (Table 5). The Evaluation section also has the highest normalised average degree (Table 4) as the clusters of authors are cited in contexts such as in Figure 9. These authors serve as 'validators' of new results being published in the field. Citations in the

Discussion section follow patterns similar to Background. These sections' main discursive function is to motivate, set-up, interpret, deepen or expand the results of the study. Highly-cited authors in this section better fit the profile of a subdomain expert with 'narrower and deeper' knowledge for their specific subdomains. Finally, similar to Evaluation, in the Conclusion section, researchers cite other studies mostly to compare the cited results to their own. However, the context semantic similarity of the Conclusion section is not as high as that of the Evaluation section.. The citation contexts in the Conclusion section have more variety than those in the Evaluation section.

The Abstract is similar to Conclusion as they both contain summaries of the paper. However, citations in Abstract do not show the aforementioned trait observed in the Conclusion section. They occur for various reasons since the Abstract being a summary of the entire paper could have citations that correspond to any section of the paper. The number of authors in Abstract in Table 3 is comparatively small, implying that authors remarkably cited in Abstract are also cited for many other purposes at various other locations. Furthermore, the Abstract section clusters into 7 (Table 4), different communities indicating the large variety in its citation contexts.

Conclusion

We conduct location-sensitive citation analysis to scrutinize the characteristics of highly-cited authors at a fine-grained level, at scale, using automated techniques from natural language processing (NLP). In particular, we sought for distinct sets of highly-cited authors in different citation locations and constructed different location specific author citation networks. We deployed several NLP tools back on the domain's own scholarly publications to analyze the literature's citation-location patterns.

Based on the results of this study, we conclude that highly-cited authors in the selected domain, CL, are distributed unevenly and display considerable variation in their citation rank. The variance is large even among highly-cited authors. Our study validates that citations usually occur towards the front portion of the article, with 37% of citations occurring in the Introduction section. However due to the characteristics of CL/NLP domain, the Method section is cited most often. In addition, for top-cited authors, citations towards the later portions of the document are significantly more prominent; specifically, they occur in Method and Evaluation sections. Furthermore, the most frequently cited section varies as we consider a larger number of top-cited authors. Our analysis of citation context based author networks show that the sub-disciplines represented across sections are different and that the traits of highly cited authors vary according to the cited sections.

The main limitation of our study is that the results are restricted to the studied domain: computational linguistics and natural language processing. However, we can apply our NLP approaches to the domain's own scholarly publications to analyze the literature's citation-location patterns. The automated extraction of citation context allows for the reported analyses to be statistically more reliable compared to previous studies, as the previous studies of citation location were largely conducted manually with limited data. We also did not disambiguate author names, although we expect the impact of ambiguous author names on the results to be minor since we limit our results to highly-cited authors which we manually inspected. Our conclusions in this study are based on results from methods that distil macro-level semantic structures in citation contexts. Understanding micro-level citation intents needs more sophisticated semantic approaches such as those employed by (Teufel et al., 2006; Shotton, 2010).

In our future work, we plan to replicate studies similar to the current study on other domains to investigate the differences between and to characterize commonalities among various domains. Finally, since our location-citation analyses are not restricted to analyses of authors, we envision the analyses to apply equally well with various other publication metadata – affiliation, nationality, among others –

in place of top-cited authors.

Reference

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *International Conference on Weblogs and Social Media, 8*, 361-362.

Bird, S., Dale, Dorr, Gibson, Joseph, Kan, Lee, Powley, Radev, & Tan. (2008) The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the Sixth International Language Resources and Evaluation Conference*, 1-5.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 10.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, *40*(4), 284-290.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003). A comparison of string metrics for matching names and records. In *KDD workshop on data cleaning and object consolidation*, 3, 73-78.

Councill, I. G., Giles, C. L., & Kan, M.-Y. (2008). ParsCit: an Open-source CRF Reference String Parsing Package. In *LREC*, 8, 661-667.

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.

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.

Flank, S. (1998). A layered approach to NLP-based information retrieval. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics,* 1, 397-403.

Garzone, M., & Mercer, R. E. (2000). Towards an automated citation classifier. In *Advances in Artificial Intelligence*. 337-346.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*. 122(3159), 108-111.

Garfield, E. (1970). Citation indexing, historio-bibliography and the sociology of science. *Proceedings of the Third International Congress of Medical Librarianship,* 187-204.

Gipp, B., & Beel, J. (2009). Citation Proximity Analysis (CPA)-A new approach for identifying related work based on Co-Citation Analysis. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics*, 2, 571-575.

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*, 29(6), 308-310.

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics*, 7(4), 887-896.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406), 414-420.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics*, 8(1), 197-211.

Joseph, M. T., & Radev, D. R. (2007). Citation analysis, centrality, and the ACL Anthology. *Technical Report CSE-TR-535-07*, University of Michigan, Department or Electrical Engineering and Computer Science.

Lipetz, B. A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation*, 16(2), 81-90.

Luong, M. T., Nguyen T. D. & Kan, M.-Y. (2010) Logical Structure Recovery in Scholarly Articles with Rich Document Features. *International Journal of Digital Library Systems*, 1(4), 1-23.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.

McCain, K. W., & Salvucci, L. J. (2006). How influential is Brooks' law? A longitudinal citation context analysis of Frederick Brooks' The Mythical Man-Month. *Journal of Information Science*, 32(3), 277-295.

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

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social studies of science*, *5*(1), 86-92.

Moravcsik, M. J., Murugesan, P., & Shearer, E. (1976). An analysis of citation patterns in Indian Physics. *Science and Culture*, *42*(6), 295-301.

Murugesan, P., & Moravcsik, M. J. (1978). Variation of the nature of citation measures with journals and scientific specialties. *Journal of the American Society for Information Science*, *29*(3), 141-147.

Nakatani, S. (2010). Language Detection Library, Cybozu Labs, Inc, Tokyo, Japan, 2010. https://github.com/shuyo/language-detection.

Porter, M. F. (1980). An algorithm for suffix *stripping*. *Program*, 14(3), 130-137.

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

Radev, D. R., Joseph, M. T., Gibson, B., & Muthukrishnan, P. (2009). A bibliometric and network analysis of the field of computational linguistics. *Journal of the American Society for Information Science and Technology*, 1001, 48109-1092.

Siddharthan, A., & Teufel, S. (2007). Whose Idea Was This, and Why Does it Matter? Attributing Scientific Work to Citations. In *HLT-NAACL*, 316-323.

Shearer, E., & Moravscik, M. J. (1979). Citation patterns in little science and big science. *Scientometrics*, 1(5-6), 463-474.

Shotton, D. (2010). CiTO, the citation typing ontology. *Journal of biomedical semantics*, *1*(1), 1.

Small, H. G. (1978). Cited documents as concept symbols. *Social studies of science*, *8*(3), 327-340.

Small, H. G. (1982). Citation context analysis, *Progress in communication sciences*, 3, 287–310.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, *87*(2), 373-388.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 103-110.

Weitz, B., & Schäfer, U. (2012). A Graphical Citation Browser for the ACL Anthology. In *LREC*, 1718-1722.

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.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2), 408-427.