

First Author Advantage: Citation Labeling in Research *

Graham Cormode
University of Warwick
G.Cormode@warwick.ac.uk

S. Muthukrishnan
Rutgers University
muthu@cs.rutgers.edu

Jinyun Yan
Rutgers University
jinyuny@cs.rutgers.edu

ABSTRACT

Citations among research papers, and the networks they form, are the primary object of study in scientometrics. The act of making a citation reflects the citer's knowledge of the related literature, and of the work being cited. We aim to gain insight into this process by studying *citation keys*: user-generated labels to identify a cited work. Our main observation is that the first listed author is disproportionately represented in such labels, implying a strong mental bias towards the first author.

1. INTRODUCTION

The notion of a citation – the formalized reference to a prior work – is at the heart of academic writing. No piece of work is complete without reference to related efforts, to set the new contribution in context. Consequently, the study of citations is a central component of understanding the relation between different articles. Indeed, the primary basis by which the impact of a piece of work is assessed is by counting the number of citations that it has received. A large number of metrics for determining the importance of a researcher, which rely on tracing citations in one way or another: *h*-index [6], *g*-index [4], and many more.

There are a broad range of studies on citations. Besides using citations to measure the impact of a paper and the influence of an author, researchers build graphs of citations [5], and study the structure, dynamics, and collaboration within and between academic fields. Leydesdorff and Amsterdamska [8] analyzed the motivation behind a citation based on surveys to authors, and examined whether the cited and citing authors are in a professional relation, whether the citation behavior is for social or for cognitive purpose.

In this work, we take a look at the process of citation from a different perspective. We focus on the process of an author making a citation, and ask what we can learn about this act. To gain a vantage point on this process, we take advantage of the fact that many researchers make use of computerized document preparation systems. In particular, systems such as Bibtex and Endnote facilitate the insertion of citations into documents. To refer to a particular work, the researcher must create a ‘key’ for it. We identify these

citation keys as objects of interest. We argue that the researchers’ choice of citation key gives us an insight into how they think about the work that they are citing. Moreover, we claim that key affects how they remember the work: if the key includes one name out of several authors, we believe that this is the name that the researcher most associates with the work.

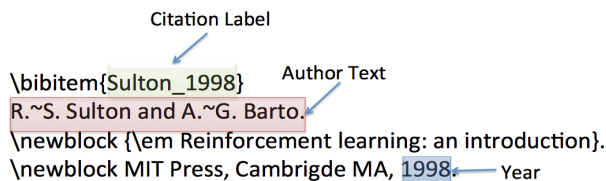
Due to increasingly collaborative research, there are many disputes regarding the order of authors [3], because the order of authors reflect the credit for contribution and authorship of the paper. Teja *et al.* [11] showed several conventions of author ordering, such as ranking authors by contribution levels and arranging each group by alphabetical order; ranking authors strictly by credit which declines with the position of authors; placing important authors in the first and last position. Our work unveils the hidden citation keys, which to some extent reflect how the researcher making a citation thinks about the contribution of authors in a cited paper.

To study the act of citation, we analyze a large data set of L^AT_EX documents and their associated bibliographies. From these, we extract titles, years and names of authors in the cited works, and measure how they relate to citation keys. In total, we identify over 506K authors referenced in 225K citations within 12K papers. We make a number of observations:

- Most strikingly the first (listed) author of the paper is very commonly included in the citation key. We argue that this gives a strong advantage to first authors, since this creates a strong link between the first author and the paper. In particular, for areas which follow rules for author ordering (e.g. alphabetical order), it can give strong benefits to authors who are often listed first.
- Other authors are not neglected: we show that subsequent authors are often referenced in the citation key, but less frequently and less prominently. A common case is to list the first author by name, and subsequent authors by initials.
- We analyze the connection between citation keys and authors given context as author ordering, time and individual habit. We show that citing authors are less likely to favor first author if cited authors are in alphabetical order. We also observe a slightly declining ratio of using authors in citation keys over time. Only a small portion of individuals stick to one labeling pattern when making many citations in a paper.
- We study other frequently occurring terms in citation keys, and observe that these include a variety of concepts: keywords indicative of the paper’s content; descriptions of the type of the paper (article, thesis, book); and other meta-data such as the year of publication.

Collectively, these give new insights into the nature of citation keys, and perhaps into how researchers think about the works they are citing.

*Authors by alphabetical order



```

\bibitem{Sulton_1998}
R.~S. Sulton and A.~G. Barto.
\newblock {\em Reinforcement learning: an introduction}.
\newblock MIT Press, Cambrigde MA, 1998.

```

Figure 1: Bib Meta Extraction

2. APPROACH

We focus on citation keys in \LaTeX source files, which reveal the hidden citation keys. Our corpus consists of the 12,611 \LaTeX sources for all papers containing references from computer science category in ‘arXiv.org’ up to April 2011. More details about the dataset can be found in our prior work [2]. Unfortunately, arXiv source files do not contain structured ‘bib’ files but only have compiled unstructured ‘bbl’ files. References are either in a separate ‘.bbl’ file or at the end of the ‘.tex’ file. We only consider papers that have references.

Given unstructured bibliographies, we aim to extract citation keys, author names, title and year. It is a special case of the general Named Entity Recognition problem [9]. However there is no existing labeled data to use for sophisticated learning algorithms. Therefore, we adopt an approach which is easy-to-implement and achieves high precision and recall.

We first introduce terminologies used through the paper. The *References* are a list of citations in a paper. A *Bib Entry* is a citation in the references. A *Citation Key* is the user-defined key to label the citation. *Bib Meta* is a collection of meta information of a Bib Entry, such as, the citation key, author names, title, year, publishing organization, and so on. *Author Text* is the piece of text in a *Bib Entry* about authors. *Authors* is a list of authors identified from *Author Text*. *Year* shows the publication year of the citation.

Our approach includes several steps.

- (1) Bib entry identification: identify bib entries in references.
- (2) Bib meta segmentation: segment the bib entry into pieces of meta info.
- (3) Author recognition: find author names (first and last names) in author text.

2.1 Bib Entry Identification

In \LaTeX , references are declared by `\thebibliography` environment [1]. We first select the reference inside valid ‘thebibliography’ environment, then extract bib entries in the reference. There are multiple commands to include a citation: `\bibitem`; the `\bibitemstart` and `\bibitemend` pair; and `\BIBentry`. We extract bib entries based on the usage of above commands. In total, we extracted **305,949** bib entries, on average each paper has **24.26** citations.

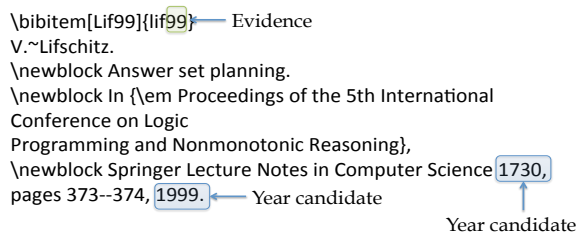
2.2 Bib Meta Segmentation

For each bib entry, we want to extract meta data of the corresponding citation. Here we focus on the citation key, authors, and year. Figure 1 shows an example of bib entry and its bib meta. We describe steps to extract each piece of meta data in the following.

Citation Key. Commands to include citations follow the format:

```
\( command ) [ ( explicit key ) ] { ( citation key ) }
```

The “explicit key” is the index printed to identify the citation. The number of explicit keys can be none or more than one; if omitted, the compiler automatically generates the citation index. The “citation key” is invented by authors. It is not printed in the final paper, and is the unique key to match up the citation context in the paper and citation entry in the reference. We scanned the bib en-



```

\bibitem{Lif99}{lif99}
V.~Lifschitz.
\newblock Answer set planning.
\newblock In {\em Proceedings of the 5th International
Conference on Logic
Programming and Nonmonotonic Reasoning},
\newblock Springer Lecture Notes in Computer Science 1730,
pages 373--374, 1999.

```

Figure 2: Evidence based year extraction

try text to match the above patterns, and extracted **304,857** citation keys, which account for 99.6% of bib entries.

Year. Some \LaTeX commands can be used to identify publication year in the bib entry. For example, ‘bibinfo’, ‘byear’ and ‘bibyear’. If any of these commands exists, we use a regular expression to extract the year. However, only a small portion of bib entries use these commands. To identify the publication year in bib entries, we search all four-digit terms which can be valid year candidates. The challenge is that these candidates can be volume number or page number, and cause false positive. For example, in Figure 1, both “1730” and “1999” are candidates.

An evidence-based algorithm is used to improve the accuracy of year extraction. Observing that quite often citation keys contain digits about the year, we compare digits in citation keys and detected four-digit year candidates. As to the example in Figure 2, the citation key is ‘lif99’ which matches the candidate ‘1999’. We focus on two-digit or four-digit sub-strings in citation keys, and match them with year candidates and return the matched candidate. There are quite a few cases where the publication year is one year later or before the year marked in citation keys. We handle such cases by allowing a ± 1 variation of the matching. If there is no evidence in the citation label, and there is more than one year candidate, we will choose the first one. By our observation, year is usually put before publishing organization, volume and page.

Author Text. There is some prior work on identifying author text in a citation. Sarawagi *et al.* [10] used hand-tuned regular expression which exploit the pattern that single letter initials before or after a word denoting last name. However, this method only handles one type of author names in citation. The authors did not provide any results on the accuracy or coverage of the method. We saw that in fact there are many different ordering of last name, first name and initials. Taking one broadly cited author for example, we found many distinct variations on the ordering: “D. E. Knuth.”, “Donald E. Knuth”, “D. E. Knuth”, “Knuth, D.”, “D. Knuth”, “Knuth, D. E.”, “D. Knuth”, “Knuth, D.E.”, “D.E. Knuth”.

When the author name contains two words and one initial, i.e., in the form of “Donald E. Knuth”, it increases the difficulty to detect author text without missing words in authors or including extra words in title. The situation gets more complex if the bib entry has more than one author, and each of them has a different ordering of first and last names, as often occurs.

We studied a large random sample of bib entries and chose 7 patterns that can guarantee accurate extraction. These patterns accompany the \LaTeX commands ‘newblock’, ‘bibinfo’, ‘Name’, ‘author’, ‘bibsc’, and so on. Patterns are matched in a fixed order, because some patterns have overlaps. Author text extracted by pattern matching are used as the ground truth set. We then design word features and adopt machine learning technique to extract author text in the remaining citations. We consider word features that can distinguish author name and not-a-name words. Table 1 gives the detail of features we used.

Table 1: Word features, examples and intuition

Word Feature	Example Words	Intuition
starts_with_brace	{A}spect	usually in title.
ends_with_brace	Theory},	The end of title, when title is enclosed by braces.
has_internal_brace	{A}spect	Part of the word has braces.
ends_with_comma	Fischer,	Punctuation: delimiter between last name and first name, or between semantic blocks, e.g., author text and title text
ends_with_period	Y. Singer.	Punctuation: the end of a semantic block, or initials.
capital_period	M.	Initials
capital_period_dup	M.M., M. M., M.-M.	Initials including middle names.
init_capital	Improved	Either last name or the beginning of the title.
four_digit_year	2006	Year
all_alpha	analysis	Likely to be words in the middle of title.
all_digits	44	Volume number, page number or other numbers.
all_symbols	"	Braces, double quotes, single quote etc.
mixed_case	ProSys, Cesa- Bianchi	Self-defined system name, algorithm name, or author name.
all_upper	ACM	Special pronoun: self-defined system or algorithm name.
all_lower	logic	Usually word in the middle of title
internal_symbol	Finite- time, Cesa- Bianchi	Hyphen connects two words in title or in few author names.
token_length	4	The number of characters in a word
summarized_pattern	[1, 22021]	The pattern of the word
token_word_id	[1, 285901]	Words removed symbols

In Table 1, the summarized pattern of a word is used to cover the various usage of capital, non-capital and symbolic characters. We summarize a word using the following heuristics. Internal symbols like hyphen are kept and unchanged.

$$\begin{array}{ll}
 [A-Z] & \rightarrow A \\
 [A\{a\}+] & \rightarrow Aa \\
 [a-z] & \rightarrow a \\
 [\{0-9\}+] & \rightarrow d
 \end{array}$$

The token word is the word after removing symbols. The word length is the length of the token word. We label each word in author text extracted in ground truth set as instance in the “NAME” class (with label 1). All words after author text in a bib entry are instances in the “NOT-A-NAME” class (with label 0). We generated 4,673,538 instances with features, which have 942,789 “NAME”, and 3,699,852 “NOT-A-NAME” instances. Note that a small number of names are also in “NOT-A-NAME” class because some authors include editor names near title text.

We apply logistic regression on word features we selected. We split the labeled set into 7:3 train and test sets. We use 5-folder cross validation and average results over 10 rounds. We obtain **0.9276** precision, **0.9251** recall and **0.926** F1-score. The result shows that features we selected are quite useful. We use the trained classifier to detect author text in unlabeled citations. Our future plan to improve performance is to model the sequential connection between words, using graphic models like conditional random

Table 2: The conditional probability of the i th author appearing in the citation key, given the number of authors

	Y = 1	2	3	4	5
1	0.62	-	-	-	-
2	0.51	0.17	-	-	-
3	0.48	0.09	0.08	-	-
4	0.45	0.04	0.05	0.04	-
5	0.40	0.03	0.03	0.04	0.02

fields. We removed \LaTeX commands in bib entries (some curly brackets are kept), so our approach is general to be applied to bib entries obtained from other sources e.g. from OCR scan in PDF or from the web.

2.3 Authors Recognition

The last step is to extract the list of author names from detected author text. Authors are not placed by a uniform format among entries, and often vary even within the same entry. Some examples:

1. Example 1: Partee,B.H., A. ter Meulen, and R.E. Wall
2. Example 2: K. Sagonas and T. Swift and D.S. Warren.
3. Example 3: Fillmore, C.J., P. Kay, and M.C. O’Connor.

We use a combined heuristic and probabilistic method to separate author names in author text. We assume that a name can be partitioned into a first name and last name. First name can be initials or full first name. Middle names are always placed in between, thus we focus on detecting the boundary between the first name and last name groups. The probabilistic method is used to identify whether a word is last name or first name, using word features. There are cases that our method failed to find the pair of first and last names. Most of them are names with single word for organization, institute, or software, e.g. “Telelogic”, “Sun”. We set the first name to be empty and assign such words as last name. In total, we identified **506,634** authors from **225,438** entries.

3. FINDINGS

3.1 First Author Advantage

We match the citation key and last names of authors by different similarity metrics to find the connection between the authors’ last names and citation key. Let string s_c be the citation key, and s_a an author’s last name. The function $f(s_a, s_c)$ returns 1 if matched, and 0 otherwise.

Exact matches. We consider exact substring matches first, i.e. a function f so that if s_a is a substring of s_c , $f(s_a, s_c) = 1$. We find that the first author’s last name has the most exact matches, which covers 54.5% of all citation keys. As the number of authors in papers ranges from one to many, we study whether this high presence of the first author’s name is affected by the number of authors. A hypothesis to test is that with more authors, the ratio of matched first author should decrease. We compute the conditional probability of citation keys matching authors in each position, given the number of authors. We use random variable $X = \{1, 2, 3, \dots\}$ to represent the number of authors, and $Y = \{1, 2, 3, \dots\}$ represent the matched author position. The conditional probability is computed as

$$\Pr[Y = i | X = j] = \frac{\Pr[Y = i, X = j]}{\Pr[X = j]},$$

where $\Pr[X = j]$ is the probability of j authors in the citation paper, and $\Pr[Y = i, X = j]$ is the probability that citation

Table 3: Similarity Metrics $f(s_a, s_c)$

Id	Metric Name	Description
M1	Exact Match	substring(s_a, s_c)
M2	Longest Sequence Ratio	$\frac{\text{lcs}(s_a, s_c)}{ s_a }$
M3	n -Gram Jaccard Similarity	$\frac{ S_a \cap S_c }{ S_a \cup S_c }$
M4	n -Gram asymmetric Similarity	$\frac{(S_a \cup S_c)^w - (S_a \setminus S_c)^w}{(S_a \cup S_c)^w}$
M5	n -Gram Dice Coefficient	$\frac{2 S_a \cap S_c }{ S_a + S_c }$

key matches the i -th author in a paper with j authors. Table 2 shows the result for $X = 1, 2, 3, 4, 5$. The row shows the value of the number of authors X , and the column in the table shows the matched author position Y . We can see that as the number of authors increases, the ratio of matching first author last name decreases only slightly. The chance that authors in higher positions are mentioned in citation keys does not increase with the number of authors.

The result shows that the first author is dramatically more likely to be included in the citation key than any subsequent author. For a two author paper, it is three times more likely that the first author will be identified in the key. For four or five author papers, the first author is approximately ten times more likely to be identified than any one of the subsequent authors. Although the first author’s presence decreases with the number of authors, it remains high (40%). This supports the notion of “first author advantage”: the idea that the first named author is much more strongly associated with the work than subsequent authors. This seems to hold, at least as far as presence in citation key is concerned. We study this issue further in our subsequent experiments.

Approximate matches. In some cases, citation keys contain fragments of last name of authors, rather than full names. To study this further, in addition to exact match, we use several other similarity metrics to estimate whether the citation key matches with an author’s last name. Metrics we used are described in Table 3. In the table, S_a is the set of n -grams of the author string s_a , and S_c for citation key string s_c . We make use of the length of the longest common substring (lcs) between two strings, and use $|\cdot|$ notation to denote the length of a string or the size of a (multi)set.

For n -Gram based metrics, we set n to be $\min(3, |s_c|, |s_a|)$. We set the threshold between matched and un-matched to be 0.5 for all metrics. For the weighted n -Gram asymmetric similarity metric, the weight parameter w is learned from manually labeled samples, as 0.5. Each metric has some advantages and disadvantages. For example, Jaccard coefficient will lead to a false negative if the citation key is much longer than last name, i.e. when the citation key contains both author’s last name and title words.

To have a clear understanding of the performance of these metrics, we manually analyzed and labeled a random selection of 432 instances. If the human judge determined the key is based on an author’s last name, the example is labeled as positive, even for the case that only the first letter of last name is used. For example, for a paper with two authors, with last names “Ladner” and “Reif” respectively, if the citation key is “LR”, we label the key as matching both authors’ last names.

Table 4 shows the performance of each metric on the labeled data. The result shows that exact match performs well but n -Gram asymmetric similarity has the best performance. We thus use the n -Gram asymmetric similarity metric to estimate the matching between citation keys and authors in each position, and it shows **61.8%** citation keys matched the first author, **21%** for the second author,

Table 4: The Performance of Similarity Metrics

Metric ID	precision	recall	F1
M1	1	0.48	0.65
M2	0.99	0.41	0.58
M3	1	0.19	0.32
M4	0.77	0.76	0.76
M5	1	0.32	0.48

Algorithm 1 Matching Author Acronym

Input: list of last names L , citation label s_c ,
string of all last names s_A

Output: True if s_c is the acronym of s_A , false otherwise

function ISACRONYM($(s_c, s_A, L, w_A, w_a, w_s)$)

$T = \text{len}(L)$.

Let A be $(|s_c| + 1) \times (|s_A| + 1)$ matrix of zeros.

for $i = 1 \rightarrow |s_c|$ **do**

for $j = 1 \rightarrow |s_A|$ **do**

$p \leftarrow \max(A[i - 1, j], A[i, j - 1])$

$w = 0$

if $\text{lower}(s_c[i - 1]) == \text{lower}(s_A[j - 1])$ **then**

$$w = \begin{cases} w_A, & s_A[j - 1] \text{ is Capital First letter} \\ w_a, & s_A[j - 1] \text{ is Non Capital, First letter} \\ w_s, & \text{otherwise} \end{cases}$$

end if

$A[i, j] = \max(p, A[i - 1, j - 1] + w)$

end for

end for

$s \leftarrow A[|s_c|, |s_A|]$

\triangleright Score for the best match

return ($s > 0.5 * w_A * T$)

end function

10% for the third author, **4%** and **1.5%** for the fourth and fifth author respectively. Thus, even allowing approximate matches and abbreviations, there is still strong evidence for a first author advantage – all positions increase their likelihood of matching compared to seeking exact matches, but the first author is still much more likely to be referenced in the citation key.

3.2 Author Acronyms

The above metrics measure whether the last name of the author in each position is used in the citation labels. However, there are cases which these will miss, such as where the key uses the acronym of last names of all authors as the citation key, e.g. “CMY”; or uses the last name of the first author and initials of the rest, e.g. “Cor-modeMY”; or the first few characters of each author’s last name, e.g. “CorMutYan”. Digits and other title words might also be attached to these patterns. To detect the presence of such acronym patterns, we used a weighted longest common sub-sequence algorithm. Algorithm 1 gives the details of the matching.

We define s_c and s_A be the citation key and the string concatenating the last names of all authors, respectively. We modify the longest common sub-sequence algorithm by assigning weights to letters in different position of the author string. The final score is then used to decide if the citation label is the acronym of all authors. We assign scores in the following way. If the matched letter is the first letter of an author’s last name, and the letter is a capital letter, we assign $w_A = 2$; if it is not a capital letter, $w_a = 1.1$. We assign $w_s = 0.1$ to other matched letters, i.e. non-initial letters of last names. The score s of a match is the sum of weighted score of matched letters. We test the score against the threshold $s > 0.5 * w_A * T$ where T is the number of authors and 0.5 is the

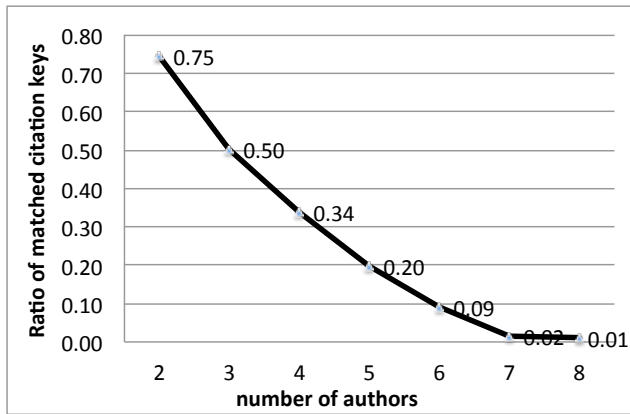


Figure 3: Ratio of author acronyms, given the number of authors

minimum threshold if lowercase acronym is matched. If the score is larger than the threshold, the matching function returns *True*, otherwise it returns *False*.

By this algorithm, we found **72%** of the citation keys contain an acronym of last names of all authors. Note that our method will typically match the case that one or more full last names are in the citation key. When we exclude the single-author cited papers the ratio of author acronym usage is **56%**. The matching algorithm reported that at least half of citation keys covered all authors, even though usually the full last name of authors in later positions is not included. This still leaves almost half of citation labels that have some different meaning and format, which are interesting to explore.

Figure 3 shows the ratio of citation keys which are an acronym of all authors as the number of authors increases. We observe a strictly decreasing line which shows that with more authors, the likelihood of using acronym of all authors reduces. However, comparing to the conditional probability of exact matching an author in high position, the coverage of all authors is much higher. From this we conclude that as there are more authors on a paper, the chance of each author getting referenced in the citation label (and so, we conjecture, figuring highly in the thoughts of the researcher making the citation) becomes lower. Moreover, the chance of being referenced falls with position: we still observe a much greater chance for the first ordered author to be referenced than any subsequent author, even when we consider acronyms.

3.3 Labeling Behavior

We analyzed the connection between the position of authors and citation keys. Here we include some context of labeling behavior, in particular, the order of authors in cited papers and the time. We also examine whether creators of citations follow a fixed pattern.

Author Ordering. We analyze whether the order of authors in citations affects citation keys. When authors are not listed in alphabetic order, it is common to rank authors by their contribution, so the first author has strongest ownership of the paper. When authors are placed in alphabetic order, chances are that the main contributor is not in the first position. Table 5 shows the breakdown of cases based on the ordering of authors (alphabetic or non-alphabetic), then by (1) whether the citation key references all authors; and (2) whether the first author’s name is given in full. Note that the ratio is computed by number of citation keys that matched the criteria to the total number of citation keys, and only citations with more than one author are considered. For (1), we see that when authors are

Table 5: Author ordering and citation keys, at least 2 authors

	author acronym match = True	author acronym match = False
in alphabetical order	43%	19.6%
not in alphabetical order	18%	19.5%
	first author match = True	first author match = False
in alphabetical order	29.5%	33.2%
not in alphabetical order	21%	16%

Table 6: Labeling pattern over time

	(1990, 2000)	(2000, 2010)
number of citations	75568	123188
number of author acronym matches	54%	46.1%
number of first author matches	40.8%	37.5%

listed alphabetically, the citation key are twice as likely to reference all authors. But when authors are not listed alphabetically, there is no great difference between the cases. For (2), when the authors are in alphabetical order, the key is more likely to not invoke the first author in full. But when not in alphabetical order, the key is more likely to invoke the first author.

The results in Table 5 supports the idea that when alphabetic order is used, people are aware that the first author is not necessarily the main contributor and thus are more likely to touch on all authors, and less likely to explicitly mention them.

Trend over Time. We conduct preliminary analysis on whether the labeling pattern changes over time. We select papers published across two decades: (1990, 2000) and (2000, 2010), and compute the ratio of citation keys with author acronym pattern and first author exact matching pattern. Table 6 shows the trend. More citations are made between (2000, 2010), in part because of the increasing number of papers published over time [7]. Interestingly, we observe an appreciable declining trend for both labeling patterns. This implies that over time, people are less commonly using author names in citation keys. Our next step will be to gather more data across time and examine the trend over longer time periods, and explore other patterns of citation keys.

Consistency. We next consider the consistency of formation of citation labels: when an author writes a paper, will he follow the same pattern to invent citation keys for all citations? We examine two patterns: exact matching the first author’s last name, and approximately matching an acronym of authors. We compute the pattern matching ratio of the paper p_i by

$$\text{pmr}(p_i) = \frac{f(\text{citations with keys matching pattern})}{f(\text{citations})}$$

If $\text{pmr}(p_i) = 1$, the authors consistently follow one pattern across all citations in the paper. We found **20.4%** of papers consistently use author acronym pattern, and **12.6%** of papers follow exact matching first author’s last name pattern. If we set the consistency ratio to 0.9, there are **32%** and **20.8%** of papers “mostly consistently” using author acronym pattern and first author last name respectively. These ratios are low, indicating that authors use various methods to compose citation keys. A conjecture is that when papers are written by multiple authors, variations in citation keys are introduced by coauthors’ different habits, and the lack of incentive to make them consistent.

3.4 n-Gram Analysis

The above results show the last name(s) of author(s) are present in some form or another in a majority of citation keys. However, we still have a large portion of citation keys that are not related to

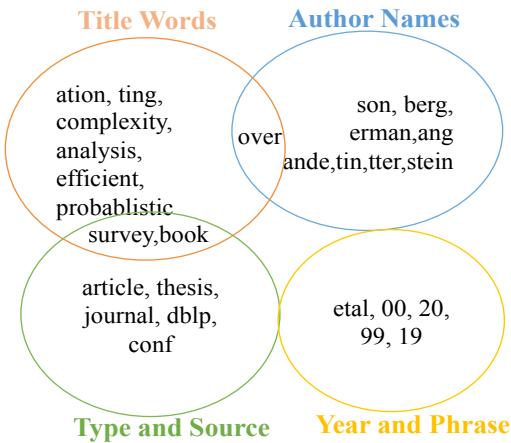


Figure 4: n -Grams Meaning Clusters

author names. To figure out the meaning hidden in a citation key, we adopt n -Gram analysis. We compute contiguous sequences of n characters from 304,857 citation keys. Here n ranges from 2 to 10, and all citation keys are lowercased.

We first observe that digits representing years are very frequent in citation keys. We compute the most frequent n -Grams and automatically cluster them based on their intersection with the extracted metadata. Figure 4 shows top-4 most frequent meanings and some of the most frequent n -grams in each group. Note that a frequent n -gram implies its $(n - 1)$ -grams are also frequent. For compactness, we only show some top 20 frequent n -grams with long length. We place these terms into four meaning clusters by computing the affinity of the term to each cluster. For example, the term ‘ation’ is in title words with probability 0.74, which is the highest affinity to clusters. Thus it is in Title Words cluster.

There are some terms that have very close affinity scores to two clusters, like “over”, which is present in both author names and title words. One occurrence in author names is “Cover1991”, and in title words is “zhu:coverage”. Similarly, the term “survey” is common in both “type and resource” and “title words”, since many survey papers use titles as “A survey of ...”. A future work is to figure out the ratio of each meaning in citation keys.

Use of Bibliographic Resources. There are many available sources of bibliographic information. For example, many journal websites allow the user to export a reference in Bibtext format, for use in their own papers. We found an interesting copying behavior due to the frequent occurrence of the term “dblp”. Some authors directly copy the bib entry from DBLP, which labels citations with its fixed format. Around 0.5% of all citation keys in our dataset are copies from DBLP. The small ratio still indicates that there is a significant amount of references taken from DBLP. We conjecture that other researchers also copy from DBLP, but modify the citation keys to fit their own habits, which reduces the value of the observed ratio. For papers with DBLP copies, around 4% of them contain more than half bib entries copied from DBLP. We also find such copying behavior more common in recent decade. **70%** of DBLP citation keys belong to papers later than the year 2000. For papers having DBLP copied citation keys, **86%** of them are later than the year 2000. We notice recently many researchers copy bibliography from Google scholar, which labels papers by last name of first name, year and first word in the title. It will be interesting to examine such behavior in a future work, and discover whether citation keys will converge.

We also observe many conference abbreviations in citation keys. However, the amount of keys citing papers in one particular conference is not large enough to draw any strong conclusions. Future work could be to match citation keys with a predefined conference list, to identify how many keys refer to the conference and what is the distribution of keys in conferences.

4. CONCLUDING REMARKS

The use of citation keys offers a rare insight into the process of making a citation, and gives some perspective into how the researcher making the citation thinks of the work being referenced. We have seen that there is a dramatically strong occurrence of the first (listed) author in such keys, far more so than other authors. We conjecture that this indicates that the first author receives much greater prominence than other authors. This “first author advantage” may have many consequences, particularly in disciplines which follow a rule (such as alphabetical ordering of authors) that mean certain researchers have a much higher chance of being listed first.

There are many further questions to address around questions of citation and citation labeling. We’ve analyzed how citation keys relate to authors in each position. It will be interesting to investigate reasons why people sometimes pick authors that are not listed first. We have indicated some ways in which citation labels are formed (from authors, from topics, from venues and from years), but it remains to fully understand these different sources, and to study what impact these have on how the work is thought of. A direction for further work is to study the citation label in the context of the citation: if we analyze the text of the paper where the citation is made, can we determine if the sentiment towards the cited work is positive or negative? Does the label give further insight into the researcher’s feelings towards the cited work’s importance or depth?

References

- [1] http://en.wikibooks.org/wiki/LaTeX/Bibliography_Management.
- [2] G. Cormode, S. Muthukrishnan, and J. Yan. *Scienceography: the study of how science is written*. the Sixth International conference on Fun with Algorithms, 2012.
- [3] A. Dance. *Authorship: Who’s on first?* Nature, 2012.
- [4] L. Egghe. An improvement of the h-index: The g-index. *ISSI newsletter*, 2(1):8–9, 2006.
- [5] N. Gilbert. A simulation of the structure of academic science. 1997.
- [6] J. E. Hirsch. *An index to quantify an individual’s scientific research output*, volume 102. National Academy of Sciences, 2005.
- [7] P. O. Larsen and M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84:575–603, 2010.
- [8] L. Leydesdorff and O. Amsterdamska. Dimensions of citation analysis. *Science, Technology & Human Values*, 15(3):305–335, 1990.
- [9] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.
- [10] S. Sarawagi, V. V. Vydiswaran, S. Srinivasan, and K. Bhudhia. Resolving citations in a paper repository. *ACM SIGKDD Explorations Newsletter*, 5(2):156–157, 2003.
- [11] T. Tschardtke, M. E. Hochberg, T. A. Rand, V. H. Resh, and J. Krauss. *Author Sequence and Credit for Contributions in Multiauthored Publications*. PLoS Biol, 2007.