DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.2, MAY. 2021

Algorithms mention in full-text content of article from NLP domain: A comparative analysis between English and Chinese

Chengzhi Zhang*, Ruiyi Ding, Yuzhuo Wang

Department of Information Management, Nanjing University of Science and Technology, Nanjing, China

ABSTRACT

Algorithms play an increasingly important role in scientific work, especially in data-driven research. Investigating the mention of algorithms in full-text paper helps us understand the use and development of algorithms in a specific domain. Current research on the mention of algorithms is limited to the academic papers in one language, which is hard to comprehensively investigate the use of algorithms. For example, in papers of Chinese conference, is the mention of algorithms consistent with it in English conference papers? In order to answer this question, this paper takes NLP as an example, and compares the mention frequency, mention location and mention time of the top10 data-mining algorithms between the papers of the famous international conference, Annual Meeting of the Association for Computational Linguistics (ACL), and the Chinese conference, China National Conference on Computational Linguistics (CCL). The results show that compared with ACL, the mention frequency of top10 data-mining algorithms in CCL is slightly lower and the mention time is slightly delayed, while the distribution of mention location is similar. This study can provide a reference for the research related to the mention, citation and evaluation of knowledge entities.

KEYWORDS

Knowledge entity；Mention of algorithms; Full-text analysis；Algorithm evaluation；Scholarly big data

1 Introduction

Influenced by the environment of big data, algorithms are widely used in scientific research and application. In 2012, the Digital Humanities Laboratory at the Swiss Federal Institute of Technology launched a project called “*Venice Time Machine*”, which used machine learning algorithms to reproduce Venice's long-standing history in a dynamic digital form (Abbott, 2017). At the same time, in most of the data-driven research, algorithms are used to process data and solve various tasks such as classification, clustering, etc.

Ding et al. (2013) proposed the *Entity Metrics*, dividing the academic entities into evaluation entities and knowledge entities, and the algorithm is a typical kind of knowledge entity. Academic papers are gorgeous sources of identifying and evaluating knowledge entities.

* Corresponding Author: zhangcz@njust.edu.cn

With more full-text database opening for free, acquiring full-text data of academic papers becomes more convenient, which provides scholars opportunities in analyzing mention or citation and evaluating academic influence of knowledge entities based on the full-text content（Belter, 2014; Pan et al., 2018; Pan et al., 2016; Wang et al., 2016）. However, researchers only consider the full-text papers in one language, which is difficult to reveal the application of knowledge entities in different countries. Meanwhile, in most research, scholars only consider the number of times that knowledge entities is cited or mentioned which might lead to a one-sided result.

Compared with other knowledge entities, algorithms did not get enough attention. Therefore, this paper takes the domain of NLP as an example, and explores the mention of the top10 data-mining algorithms（Wu et al., 2008）in the full-text academic papers respectively. Specifically, this article attempts to explore the *mention frequency, mention location and mention time* of the algorithms in NLP papers. NLP is a domain centering on data processing and technical practice, in which data-mining algorithms are extensively used. Therefore, academic papers of NLP are suitable for exploring mention of data mining algorithms. It is believed that the comparative study helps scholars understand the usage and distinction of algorithms in a specific domain, and also enables them to comprehend the differences and similarities in the research of different areas.

2 Related works

Knowledge entities are the mediums of knowledge units in the scientific literature, including keywords, datasets, algorithms, software, key methods, theories and so on（Ding et al., 2013）. At present, the research on the mention or citation of knowledge entities mainly focus on datasets and software.

A summary of the research on mention of dataset: Currently, the application of dataset in academic research is not standardized（Samiya et al., 2017）. In order to guide quoting dataset scientifically and promote dataset sharing, many scholars began to study the mention or citation of dataset. Belter and Browman（2014）took three datasets in the field of oceanography as the objects and counted the cited times so as to evaluate them. Wang et al.（2016）collected full-text papers of bioinformatics and evaluated the influence of dataset according to the cited and downloaded times. Robinson-Garcia et al.（2016）found that there are large differences in citation of dataset among different disciplines.

A summary of the research on mention of software: The research on the mention or citation of software provides insights into the rules of scholars using software and reveals the role of software in scientific research（Pan, 2018）. Pan et al. conducted a series of studies on the mention, citation and evaluation of software in the full-text papers, and found that there are a large number of non-standard citations（Pan, 2018; Pan, 2016; Pan et al., 2015）. Howison et al. explored the way software was mentioned in the full text of academic literature in biology, discovering that most of the citation did not meet the specification（Howison, 2016）. Li (2017) considered the mention and citation of R software as well as its software packages in the academic articles, finding that core software, software packages and software functions play different roles in scientific research.

However, algorithm, as a knowledge entity（Ding et al., 2013）, got little attention on the mention or citation in previous work. Mike and Nabeil.（2015）concluded that the mention of algorithm in journal papers has increased dramatically from 2004 in Library and Information Science. Wang et al. examined the mention of algorithms in the papers of ACL conference based on the frequency, location and tasks that algorithms used to solve and evaluating the influence of them (Wang & Zhang, 2017; Wang & Zhang, 2018; Wang & Zhang 2020) .

To summarize, the existing research about the mention or citation of knowledge entities mainly focuses on dataset or software, lacking to the algorithms; in addition, most studies only consider the times of citing, leading to one-sided results.

3 Methodology

Research on NLP includes the processing of large amounts of data and it's suitable for studying the mention of data-mining algorithms. As shown in Figure 1, firstly, this study collected full-text dataset of papers from two conferences, and compiled an algorithm dictionary manually. Then we extracted algorithm sentences based on the dictionary, and examined the mention of top10 data-mining algorithms, including *mention frequency*, mention location, and mention time. Finally, we compared the obtained results between the two conferences.

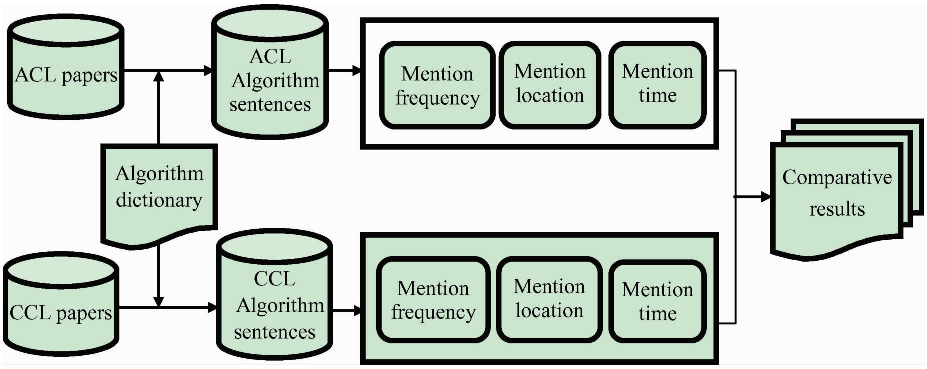


Figure 1 Framework

3.1 Dataset

In the field of computer-related disciplines, the impact of conference papers is higher than that of journal papers, and the research topics are more advanced (Qian et al., 2017; Lorcan et al., 2010) . Thence the conference papers of NLP are collected as the dataset for our research. The full-text paper of two top-tier conferences in NLP domain, namely, ACL[1] and CCL[2] were choose. ACL is the highest-level international academic conference in NLP, sponsored by the Computational Linguistics Association; CCL is the most famous and largest Chinese academic conference in NLP.

We obtained the full-text data between 1993 and 2016 from the website of the two conferences, including 4,471 ACL papers in XML format and 1,767 CCL papers in PDF format. In order to facilitate computer processing, we converted all PDF documents into plain text (TXT) format by the OCR function of CAJVIEVER[3]. Subsequently, we manually label the 1,767 full-text papers in TXT format into XML format according to the tags shown in Table 1, and

[1] ACL papers are downloaded from ACL Anthology (Digital Archives of Computational Linguistics Research Papers) at *https://aclanthology.coli.uni–saarland.de*

[2] CCL papers are downloaded from the official website of the Computational Linguistics Committee of the Chinese Information Society at *http://www.cips–cl.org/webmirror*

[3] The official website of the software is: http://cajviewer.cnki.net

corrected the format errors and scrambled letters. Figure 2(a) and 2(b) show the specific labeling result of English paper and examples of Chinese paper respectively.

Table 1 Tag description of dataset

Tag	Description
<title_chinese>... </title_chinese>	Chinese title of paper
<author_chinese>... </author_chinese>	Chinese name of author
<abstract_chinese>... </abstract_chinese>	Chinese abstract
<keyword_chinese>... </keyword_chinese>	Chinese keywords
<title_english>... </title_english>	English title of paper
<author_english>... </author_english>	English name of author
<abstract_english>... </abstract_english>	English abstract
<keyword_english>... </keyword_english>	English keywords
<chapter>... </chapter>	Chapter starts and ends
<chapter_title>... </chapter_title>	Heading of chapters
<chapter_sub_title>... </chapter_sub_title>	Heading of sections
<para>... </para>	Para starts and ends
<section type=<Introduction|Related work|Method|Evaluative|Result|Conclusion>... </section>	Type of every chapter

```
<title_english>Power Law for Text Categorization</title_english>
<author_english>Wuying Liu, Lin Wang, Mianzhu Yi</author_english>
<Abstract_english><section=abstract>Text categorization（TC）is a challenging issue, both in the TREC
email spam filtering task and the Chinese web…..</Abstract_english>
<keyword_english><section=keyword>Text Categorization, Power Law, Online Binary TC, Batch
MultiCategory TC, TREC</keyword_english>
<chapter>
<chapter_title><section=introduction>1 Introduction</chapter_title>
<para>Automated text categorization（TC）has been widely investigated since…..</para>
……
</chapter>
```

(a) A labeling example of English paper[4]

```
<title_chinese>基于量词的名词概念获取研究</title_chinese>
<author_chinese>王萌，俞士汶</author_chinese>
<abstract_chinese><section=abstract>概念获取是自然语言理解领域中重要的研究课题……可以
区分大部分名词概念。</abstract_chinese>
<keyword_chinese><section=keyword>概念获取；量名搭配；量词；聚类</keyword_chinese>
<chapter>
<chapter_title><section=introduction>1 引言</chapter_title>
<para> 概念获取（Concept Acquisition）又称概念学习（Concept Learning）……</para>
......
</chapter>
```

(b) A labeled example of Chinese paper[5]

Figure 2 Example of labeling result

[4] The original paper corresponding to this sample is at: http://www.cips–cl.org/static/anthology/CCL–2013/CCL–13–064.pdf

[5] The original paper corresponding to this sample is at: http://www.cips–cl.org/static/anthology/CCL–2014/CCL–14–008.pdf

3.2 Algorithm dictionary construction and algorithm sentences extraction

We used a dictionary-based approach to identify algorithms from the full text of academic papers. According to *Top 10 Algorithms in Data Mining* (Wu, 2008), we obtained the standard names of the ten algorithms. Then we used these full names as queries to search on Google Scholar[6] and Wikipedia[7], acquiring alias for each algorithm based on the descriptions of algorithms in related papers and Wikipedia explanations. In addition, Chinese alias of algorithms were collected with the same approach on CNKI[8] and Baidu Baike[9]. Finally, we constructed the top10 algorithm dictionary shown in Table 2.

Table 2 Top10 data-mining algorithm dictionary

No.	Standard name	Alias in English	Alias in Chinese
1	C4.5	–	-
2	K–means	k means	k 均值, k 平均
3	Support vector machines	support vector machine、svm、svms	支持向量机, 支撑向量机
4	Apriori	–	-
5	expectation maximization	expectation–maximization、EM	最大期望算法, 期望最大化算法
6	PageRank	PR	-
7	Adaboost	Adaptive Boosting	-
8	K–nearest neighbor	KNN、k–nn、k nearest neighbor、k nearest neighbour、k nearest neighbors、k nearest neighbours、k –nearest neighbors	K 最近邻, k 近邻
9	Naïve Bayes	Naïve –bayes、naïve –bayes、NB、Naive Bayesian、Naive Bayes	朴素贝叶斯
10	CART	classification and regression trees	-

For a name of an algorithm, we matched it with the content of each sentence in a paper, and the specific sentence containing the name were defined as *algorithm sentence*. The name of algorithm and algorithm sentence containing the name were recorded simultaneously. At the same time, the title of chapter where the algorithm sentence locates and the ID of article were also recorded. Finally, we acquired a total of 8,303 algorithm sentences with the corresponding type of chapter and article ID, including 5,975 algorithm sentences from 1,323 ACL papers, and 2,328 algorithm sentences from 354 CCL papers.

3.3 Comparative analysis on the mention of algorithms

(1) Mention frequency of algorithms

The mention frequency is the times that an algorithm is mentioned in a paper. We divided the mention frequency into two indicators: *the number of papers mentioning algorithms and the average times of mention*. The number of papers mentioning algorithms refers to the

[6] The website of Google Scholar is: https://scholar.google.com

[7] The website of Wikipedia is: www.wikipedia.org

[8] The website of CNKI is: https://cnki.net

[9] The website of Baidu Baike is: https://baike.baidu.com

Count One, regarding the article as the statistical unit, that is, no matter how many times the algorithm is mentioned in a paper, it is only recorded once. The average times of mention is calculated by formula（1）, in which *total times of mention* refers to the *Count X*（Ding et al., 2013a）, taking the sentence as the statistical unit, that is, the number of sentences mentioning the algorithm.

$$\textit{Average times of mention} = \frac{\textit{Total times of mentions}}{\textit{Number of papers mentioning algorithms}} \quad (1)$$

In addition, considering the differences of authors’ writing style, or detailed or abbreviated, there might be a large distinction in the descriptions of algorithms, which might have a great impact on the total times of mentions. Therefore, in the analysis of mention location and mention time, we used the number of papers mentioning algorithms as the main indicators for statistics.

(2）Mention location of algorithms

Mention location is the type of chapter where the algorithm is mentioned in a paper. Lin et al. studied the papers of empirical research and found that the most common structure was *introduction, methods, results, discussion* and *conclusion*, which was named the IMRDC（Lin, 2012）. Based on this and combined with the characteristics of NLP and the structure of academic papers in the domain, we divided the types of chapter into 7 species: *abstract, introduction, related work, method, evaluation, discussion & result, conclusion*. The annotation of mention location was completed by a master and a doctor. In order to check the consistency of annotation, we randomly selected 50 articles which were tagged by them independently, and then calculated the *Kappa* coefficient (Warrens, 2001) based on the annotation result. The *Kappa* coefficient was 0.84, which indicated the sufficient reliability of one labeler annotating all of the papers. Therefore, the master student annotated all of the remaining papers. On this basis, we counted the number of papers mentioning algorithms in the seven chapters.

(3）Mention time of algorithms

Mention time means the time at which the paper containing algorithms published. Both ACL and CCL assign a unique ID embodying the publishing year to each paper, for example, *CCL2003-71* and *ACL2014-1055* are separately published in 2003 and 2014. For each algorithm, we counted the number of papers mentioning the algorithm each year.

4 Results

4.1 Comparison on mention frequency of algorithms

We separately compare the proportion and ranking of the numbers of papers mentioning algorithms and the average times of mentions for each algorithm（Proportion means the ratio of *the number of papers mentioning algorithms to the total number of papers*）. The results are shown in Figures 3 & 4 and Tables 3 & 4.

As shown in Figure 3 and Table 3, in ACL and CCL, the number of papers mentioning “*SVM*” is the highest, which is even more than 50% in CCL.The result indicates that in NLP domain, whether it is a specific Chinese conference or an international conference, the most frequently mentioned algorithms is *SVM*, reflecting the scholar's preference for it and its importance for NLP task. Since *SVM* has a solid theoretical foundation and is one of the most stable and accurate algorithms among all-known algorithms（Wu et al., 2008）. In contrast, the proportions of *Apriori* and *CART* are both low, which reveals that in NLP domain, *Apriori* or *CART* is neither widely used. *EM* has the largest disparity in the number of papers mentioning algorithm between ACL and CCL, ranking second in the ACL, but only 2.9% of papers in CCL mentioning the algorithm. The same large gap also appeared in the influence of *KNN.* This might be caused by the different research emphasis of the two conferences, since CCL pays more attention on Chinese natural language processing.

Figure 4 and Table 4 show the comparison about the average times of mention in ACL and CCL. Seven algorithms in CCL get higher average mention than that in ACL, which reveals that the explanation and description of the algorithm may be more detailed in CCL papers. We speculate that Chinese scholars may pay more attention to the introduction and interpretation of the algorithms. The algorithms with large gaps in the two conferences *is Adaboost. Adaboost* ranks first in the average times of mentions in CCL, which is the highest among all algorithms. However, it only ranks eighth in ACL. At the same time, *Adaboost* get the lower number of papers mentioning algorithms. In summary, we speculate that scholars might prefer to describe a less commonly used algorithm more in detail.

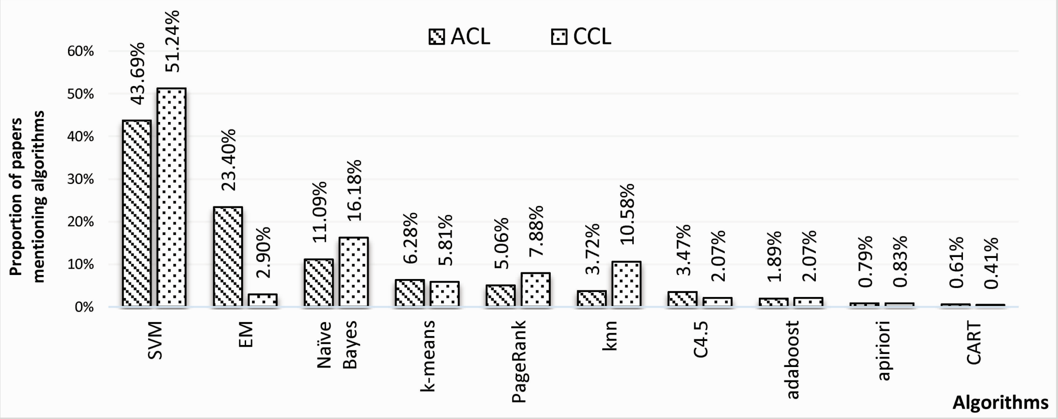


Figure 3 The proportion of papers mentioning algorithms in the two conferences

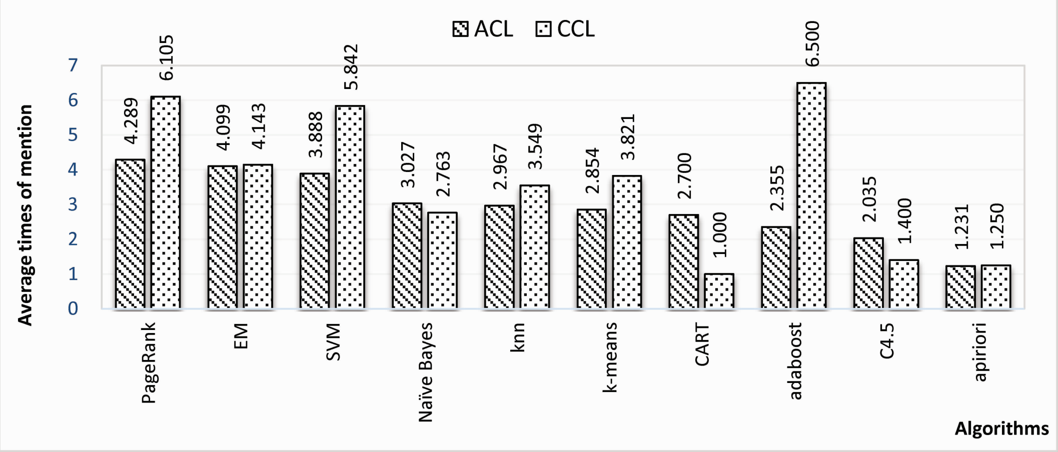


Figure 4 The average times of mention in the two conferences

Table 3 The number and proportion of papers mentioning algorithm in the two conferences

Info. / Rank	ACL		CCL	
	Name	The number of papers /proportion	Name	The number of papers /proportion
1	SVM	717 / 43.69%	SVM	247 / 51.2%
2	EM	384 / 23.40%	Naïve bayes	78 / 16.18%
3	Naïve bayes	182 / 11.09%	KNN	51 / 10.58%
4	K–means	103 /6.28%	PageRank	38 / 7.88%
5	PageRank	83 / 5.06%	K–means	28 / 5.81%
6	KNN	61 / 3.72%	EM	14 / 2.90%
7	C4.5	57 / 3.47%	C4.5	10 / 2.07%
8	Adaboost	31 / 1.89%	Adaboost	10 / 2.07%
9	Apiriori	13 / 0.79%	Apiriori	4 / 0.83%
10	CART	10 / 0.61%	CART	2 / 0.41%

Table 4 The average times and ranking of mention in the two conferences

Info. / Rank	ACL		CCL	
	Name	Average times	Name	Average times
1	PageRank	4.289	Adaboost	6.500
2	EM	4.099	PageRank	6.105
3	SVM	3.888	SVM	5.842
4	Naïve Bayes	3.027	EM	4.143
5	KNN	2.967	k–means	3.821
6	k–means	2.854	KNN	3.549
7	CART	2.700	Naïve Bayes	2.763
8	Adaboost	2.355	C4.5	1.400
9	C4.5	2.035	Apiriori	1.250
10	Apiriori	1.231	CART	1.000

The results above show that in a specific domain, even the general algorithm, the mention frequency is not same, and on the contrary it will show a big difference in some aspects. We think the difference may be mainly caused by the overall research topic, writing style or so on. These findings may give a reference to scholars who are studying writing style or more.

4.2 Comparison on mention location of algorithms

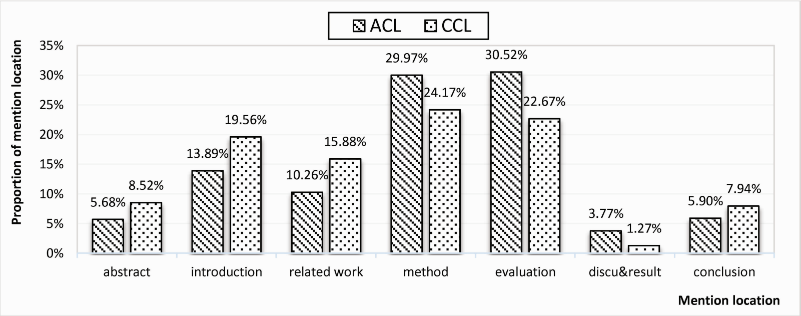


Figure 5 The proportion of mention location in two conferences

As displayed in Figure 5, *proportion* means the ratio of *the number of papers mentioning algorithms* in each type of chapter to *the total number of papers*. Considering there exists bilingual abstracts in CCL papers, we only examine the Chinese abstract for CCL papers. In the two conferences, the top 10 data-mining algorithms are mentioned most in method of CCL and evaluation of ACL, both reaching over 20%. Next is introduction, then related work; in abstract and conclusion, the mentions are much less, and the least is discu&result. Comparing the two conferences, we find that the mentions of algorithms in method and evaluation of ACL are higher than that of CCL. At the same time, the mentions of algorithms in introduction and related work of the CCL paper are higher than that of ACL. In this regard, we infer that the paper accepted by ACL pay more attention to the innovation and breakthrough of algorithm, and the paper accepted by CCL more concerned about the use of algorithms by predecessors.

Table 5 shows the similar ranking of the top10 data-mining algorithms in the two conferences. As method and evaluation are both related to approaches used by scholars, we repute that algorithms mentioned in these two sections could be regarded as actually-used ones in a paper. Therefore, we take *method* and *evaluation* as the objects, and further investigate the added proportions of each algorithm in these two sections. It can be seen from the previous results that the mention frequency of *Apriori* and *CART* are too low in both ACL and CCL. In order to avoid contingency, we eliminate these two algorithms and study on the other eight.

Table 5 The ranking of mention location in two conferences

Rank	1	2	3	4	5	6	7
ACL	evaluation	method	introduction	related work	conclusion	abstract	discussion & result
CCL	method	evaluation	introduction	related work	abstract	conclusion	Discussion & result

Table 6 The added proportion of each algorithm in method and evaluation (%)

Algorithm	SVM	Naïve Bayes	KNN	PageRank	k-means	EM	Adaboost	C4.5
ACL	64.95	67.98	73.81	61.70	74.82	56.39	67.50	84.00
CCL	44.21	43.55	53.25	52.31	68.89	54.55	45.45	75.00

Table 6 indicates the added proportion of eight algorithms in *method* and *evaluation.* It can be seen that the added proportion of each algorithm in the two sections is commonly high, and every added proportion in ACL is more than 50%, among which the highest is C4. 5, accounting for 84%, and in CCL C4.5 is also the highest. The added proportion of each algorithm in CCL exceeds 40%. In summary, the mentions of algorithms are mostly in the *method* and *evaluation*, further illustrating that in the domain of NLP, the top10 data-mining algorithms are primarily used as concrete experimental approaches.

4.3 Comparison on mention time of algorithms

(1) The proportion of papers mentioning algorithms

The proportion of papers mentioning algorithms each year is displayed in Figure 6. Between ACL and CCL papers, the number of papers mentioning algorithms both shows a rising with fluctuations. In ACL, the mention of the top10 algorithms began in 1993, but in CCL it appeared in 2001, indicating that compared with international conference, related research in Chinese conference has started later. However, it can be seen from the overall trend that the fluctuations of the number of papers mentioning algorithms in the two conferences is similar, both existing a large increasing in 2005-2006 and 2012-2013, and an apparent decreasing in 2012 and 2016. Although the mention of top10 algorithms in CCL appears later, the overall trend is similar to that of ACL. Ii is obvious the gap between Chinese research and frontier international research is getting smaller and smaller, and the speed of following in Chinese research is getting faster.

In addition, we also explore the evolution of each algorithm mentioned in the two conferences. Similarly, we study on eight algorithms expect *Apriori* and *CART* cause the mention of them are too low in both ACL and CCL. Figure 7 shows the *proportion* of papers mentioning algorithms respectively in ACL and CCL from 1993 to 2016.

Compared with ACL, the mention of each algorithm in CCL did not appear until 2000. The overall changing of *SVM* and *Naïve Bayes* is similar, both showing a gradual increase with small twists and turns. As the most frequently mentioned algorithm in the two conferences, *SVM* has solid theoretical foundation（Wu et al., 2008）so that large number of scholars use it to solve research problem, especially the classification task; Naïve Bayes algorithm is also widely used in text classification and spam filtering, etc., with simple principle and good effect（Wu et al., 2008）. Both of the two algorithms are very popular among scholars especially as a baseline in an experiment. There are no special changes of the other six algorithms from 7(c）to 7(h）, the trend of CCL is more volatile than that in ACL, and the zero mention is also more. This reflects the wider application of the six algorithms in ACL. In addition, we also discover that during 2015-2016, most of the mentions in ACL and CCL shows a downward trend. We speculate that this may be due to the emergence of better performance algorithms, such as *neural networks*, which are not in the scope of this study. Scholars have tended to use these algorithms to deal with problems.

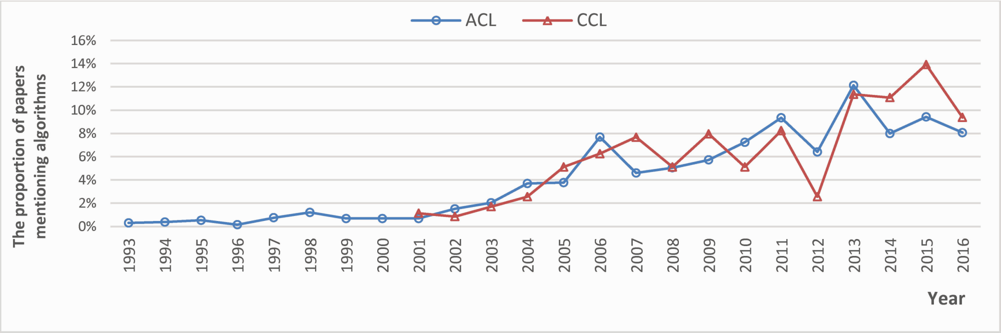
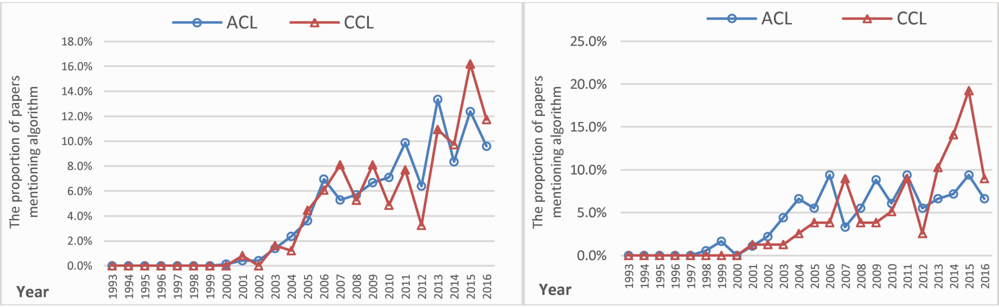
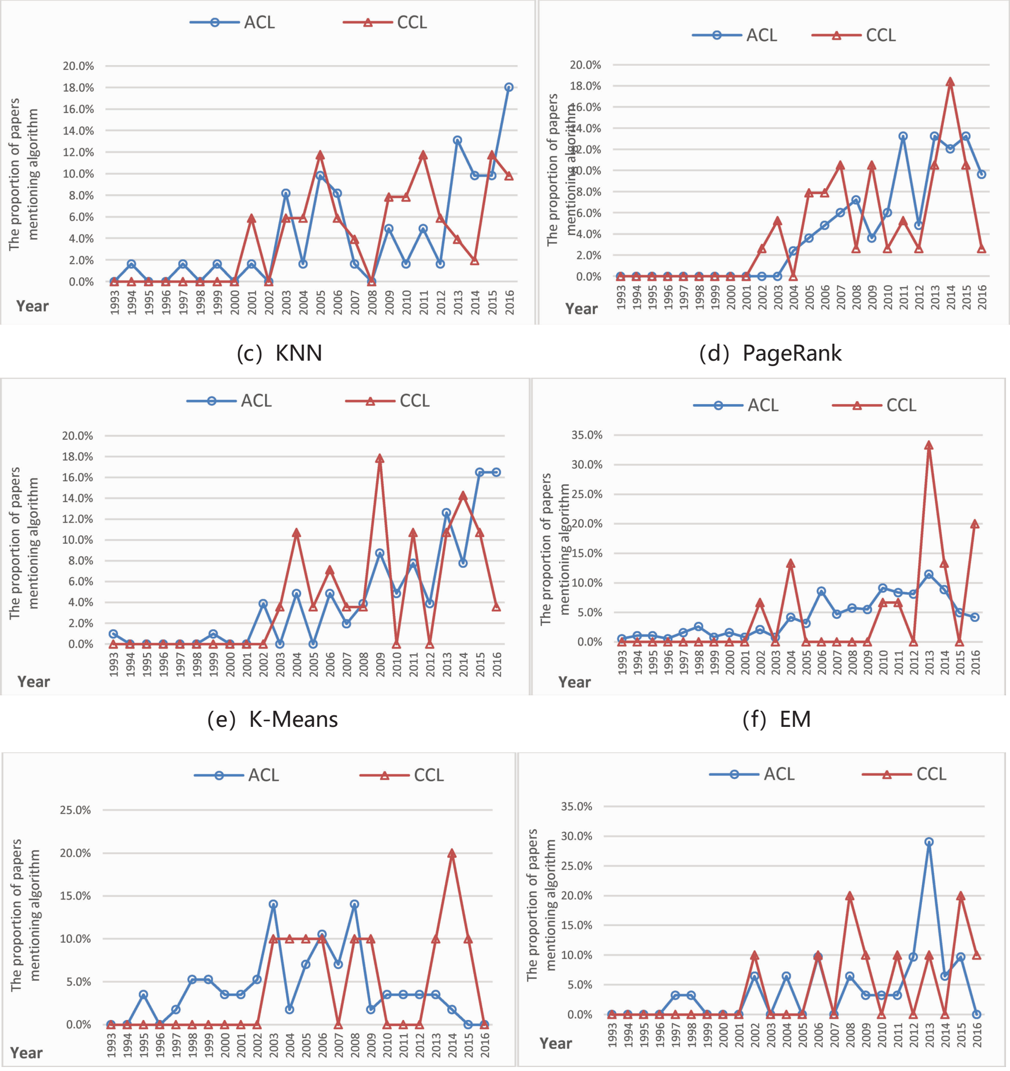


Figure 6 The proportion of papers mentioning algorithms each year



(a) SVM (b) Naïve Bayes



(c) KNN

(d) PageRank

(e) K-Means

(f) EM

(g) C4.5

(h) Adaboost

Figure 7 The proportion of papers mentioning algorithms each year

(2) The first appearance of algorithms

From the results of the mention time of the top10 algorithms, we notice there existed a lag in Chinese conference in the early stage. On this basis, we further explore the time when each algorithm first appeared in the two conference from 1993 to 2016. Defining t as the time difference for the first occurrence of each algorithm in ACL and CCL, we calculate it according to formula (2) .

$$t = t_{F\text{-}CCL} - t_{F\text{-}ACL} \tag{2}$$

Where $t_{F\text{-}CCL}$ and $t_{F\text{-}ACL}$ respectively means the first time when an algorithm mentioned in CCL and ACL. We also eliminate *CART* and *Apirior* to avoid contingency.

As Figure 8 shown, in addition to *PageRank*, the other seven algorithms mentioned in ACL are all earlier than that in CCL, and there are five algorithms having a time difference more than 5 (including 5) years, which reflects that from 1993 to 2016, the mention in Chinese

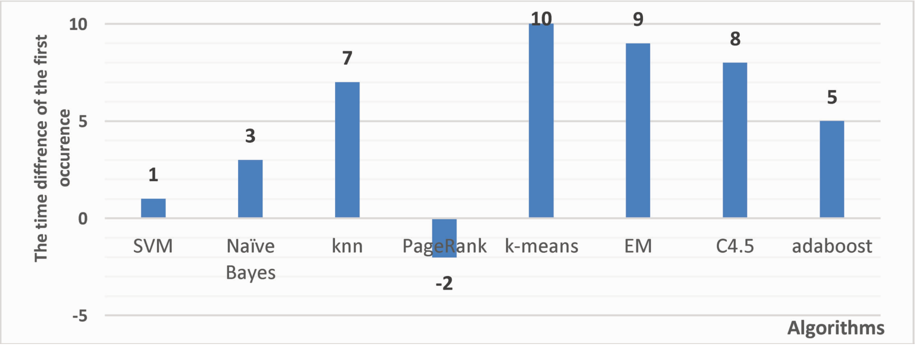


Figure 8 The time difference of the first occurrence of each algorithm (year)

conference is falling behind. Therefore, we use the content-analysis method to investigate the articles that containing the first occurrence of the eight algorithms in the ACL and CCL. The results are shown in Table 7.

In CCL, all the algorithms first appeared after 2000, and the *SVM* and *KNN* were first mentioned in the introduction section of the same article , which indicates the two algorithms were mentioned as research background, not the method. PageRank and Adaboost were also in introduction. *Naïve Bayes , K-means , EM , and C4.5* are in *method or evaluation* , actually used to deal with the problem. The results of the ACL are quite different. Firstly, six of the eight algorithms were mentioned before 2000 , only two mentioned in *introduction* and *related work* as background descriptions , and the rest were actually used as experimental approaches. To some extent , the result indicates that the mention of algorithm in ACL is earlier and research emphasized on the practice of algorithms more than that in CCL. Additionally , the problems solved by scholars using algorithms are different, which may be related to the respective characteristics of each algorithm and the differences in the language of dataset.

5 Conclusion and future work

Studying the mention of the algorithm in the full-text paper enables us to comprehend the overall application of the algorithm in a specific domain. This paper takes the NLP domain as an example, selects the top 10 data-mining algorithms as objects, and investigates the mention of algorithms in full-text papers of two different conference. The results show that in the two conference, the mention of the algorithms is different on frequency and time but similar on location. In terms of frequency, *SVM* has the highest number of papers mentioning algorithms and *CART* is the lowest both in ACL and CCL, but *EM* differs greatly between the two conferences; in terms of location, the distribution of each algorithm in the two conferences is similar, both mentioned most in method and evaluation; in terms of time, the first mention of each algorithm in CCL appears later than ACL, but as time progresses, the overall evolution trend has tended to be similar. In conclusion, this study can provide a reference for the related research about mention and citation of knowledge entities, enrich the field of full-text analysis.

Compared with the previous studies, we conduct analysis based on the full text of academic papers, not only considering the frequency, but also investigating the locations for each mention and the changes of time. The results are more comprehensive. But there are still

Table 7 Information of articles mentioning algorithms for the first time in ACL and CCL

Type / Algorithm	ACL					CCL				
	Title	Author	First mention time	First mention location	Problems solved	Article title	Author	First mention time	First mention location	Problems solved
SVM	Distribution –Based Pruning of Backoff Language Models	Jianfeng Gao	2000	Evaluation	n –gram backoff models pruning	基于改进的贝叶斯模型的中文网页分类器(A modified statistic Chinese Web Page Classifier)	Qin Bing, Zheng ShiFu, Liu Ting	2001	Introduction	Chinese Web Page Classification
Naïve Bayes	Part of Speech Tagging Using a Network of Linear Separators	Dan Roth, Dmitry Zelenko	1998	Evaluation	Part of Speech Tagging	英语句法分析树向汉语分析树的转换(English Chinese Transfer of a syntactic Parsing Tree)	Yao Jianmin, Zhang Jing, Zhao Tiejun	2001	Method	Translation Selection and Translation Generation
KNN	Similarity –Based Estimation of Word Co–occurrence Probabilities	Ido Dagan, Fernando Pereira.	1994	Introduction	Word Co –occurrence Probabilities	基于改进的贝叶斯模型的中文网页分类器(A modified statistic Chinese Web Page Classifier)	Qin Bing, Zheng ShiFu, Liu Ting	2001	Introduction	Chinese Web Page Classification
K–MEANS	Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning	Vasileios Hatzivassiloglou, Kathleen McKeown	1993	Method	Clustering Adjectives According to Meaning	话题检测与跟踪技术的发展与研究(Development and Analysis of Technology of Topic Detection and Tracking)	Luo Weihua Liu Qun Cheng Xueqi	2003	Method	Topic Detection and Tracking
EM	An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora	Julian Kupiec	1993	Method	noun phrase correspondences	Experiments on Unsupervised Chinese Word Segmentation and Classification	Jinhu Huang, David Powers	2002	Evaluation	Chinese word segmentation and classification
PageRank	Paragraph–, word–, and coherence –based approaches to sentence ranking: A comparison	Florian Wolf	2004	Method	Sentence ranking	基于遗传算法的定题信息搜索策略 (Focused Crawling Based on Genetic Algorithm)	Xu Huanqing, Wang Yongcheng, Sun Qiang	2002	Related work	Question –setting Information Search
Adaboost	Mistake –Driven Mixture of Hierarchical Tag Context Trees	Masahiko Haruno	1997	Related work	learning a tag model	基于大规模语料库的英语从句识别(English Clause Recognition Based on Large–scale Corpus)	Huang Yu, Li Sheng, Meng Yao	2002	Introduction	English clause recognition
C4.5	Combining Multiple Knowledge Sources for Discourse Segmentation	Diane Litman, Rebecca Passonneau	1995	Method	Discourse Segmentation	汉语部分分析研究(Research on Chinese Partial Parsing)	Zhou Qiang	2003	Method	Part Analysis of Chinese Language

many limitations. First of all, this study only considers the top 10 classical data-mining algorithms. In fact, there are many other commonly used algorithms in the NLP domain so in the future, we will expand the kinds of algorithms, and transform the way to extract algorithm sentence, such as combining the methods of named entity recognition. What's more, there are specific algorithms in different disciplines and are all playing an important role, in the future work we will study and evaluate these special algorithms in corresponding professional fields. Secondly, this study only considers the frequency, location, and time of mention and mainly uses frequency in the whole analysis, not pays concentrate to the semantic level of mentions such as the motivations, the tasks that the algorithms are used to solve, etc. Next, we will add the semantic information to examine the mention of algorithms more deeply. Besides, this study uses a dictionary-based matching method to identify the algorithm mentioned in the paper, but does not premeditate the issue of anaphora, that is, the author did not use the name of algorithm every time in description, but used pronouns referring to the algorithm. We will improve on this in next step. Then, this study takes the papers of ACL and CCL as the objects to analyze and compare the mention of algorithms. However, the number of conference papers is small, especially in CCL. In the future, we will consider expanding the dataset and supplementing more papers in NLP to investigate the mention of algorithms more comprehensively. Finally, in the future, we will also combine the citation of algorithms to investigate the citation frequency, the citation location, the citation motivation and the citation evolution over time.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (Grant No. 72074113) .

References

Abbott, A.(2017). The "time machine" reconstructing ancient Venice´s social networks. *Nature, 546* (7658), 341–344.

Belter, C. W., & Browman, H. I.(2014). Measuring the Value of Research Data: A Citation Analysis of Oceano–graphic Data Sets. *PLOS ONE, 9* (3), e92590.

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

Ding, Y., Song, M., Han, J., et al. (2013b). Entitymetrics–measuring the impact of entities. *PLOS ONE, 8*(8), 1–14.

Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technol–ogy, 67* (9), 2137–2155.

Li, K. (2017). How is R cited in research outputs? Structure, impacts, and citation standard. *Journal of Informet–rics, 11* (4), 989–1002.

Lorcan, C., Jill, F., Barry, Smyth., et al.(2010). A quantitative evaluation of the relative status of journal and con–ference publications in computer Science. *Communications of the ACM, 53* (11), 124–132.

Mike, T., & Nabeil, M. (2015). How important is computing technology for library and information science re–search?. *Library & Information Science Research, 37* (1), 42–50.

Pan, X. (2016). Disciplinary differences of software use and impact in scientific literature. *Scientometrics, 109* (3), 1593–1610.

Pan, X. (2018). Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A com–

parative study of three tools. *Journal of Informetrics, 12* (2), 481–493.

Pan, X., Yan, E., Wang, Q., & Hua, W.(2015). Assessing the impact of software on science: A bootstrapped learning of software entities in full–text papers. J*ournal of Informetrics, 9* (4), 860–871.

Qian, Y., Rong, W., Jiang, N., Tang, J., & Xiong, Z.(2017). Citation regression analysis of computer science publications in different ranking categories and subfields. *Scientometrics, 110* (3), 1351–1374.

Robinson–García, N., Jiménez–Contreras, E., & Torres–Salinas, D. (2016). Analyzing data citation practices us–ing the data citation index. Journal of the Association for Information Science and Technology, 67(12), 2964–2975.

Samiya, K., Liu, X., Kashish, A.S., & Alam, M.(2017). A survey on scholarly data: From big data perspective. *Information Processing & Management, 53* (4), 923–944.

Wang, X., Ma, S., Yu Z.L., et al. (2016). Research on citation behavior and influence of scientific data. *Journal of The China Society for Scientific and Technical Information, 35* (11), 1132–1139. (In Chinese)

Wang, Y., & Zhang, C. (2017). Research on academic impact of algorithms using full text content. *Library and Information Service, 61* (23), 6–14. (In Chinese)

Wang, Y., & Zhang, C. (2020). Using the full–text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of Informetrics, 14* (2): 101091.

Wang, Y., & Zhang, C. (2018). Using full–text of research articles to analyze academic impact of algorithms. In: *Proceedings of iConference 2018*, Sheffield, UK.

Warrens, M. (2001). Chance–corrected measures for 2×2 tables that coincide with weighted kappa. *British Journal of Mathematical and Statistical Psychology, 64* (2), 355–365.

Wu, X., Kumar, V., Quinlan, J.R., et al. (2008). Top 10 algorithms in data mining. *Knowledge information sys–tem, 14* (1), 1–37.