Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 1827074, 21 pages
https://doi.org/10.1155/2018/1827074

Review Article

A Bibliometric Review of Natural Language Processing Empowered Mobile Computing

Xieling Chen,[1] **Ruoyao Ding**,[2] **Kai Xu**,[3] **Shan Wang**,[4] **Tianyong Hao**,[5] **and Yi Zhou**[6]

[1]*College of Economics, Jinan University, Guangzhou, China*
[2]*School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China*
[3]*School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China*
[4]*Department of Chinese Language and Literature, University of Macau, Macau SAR, China*
[5]*School of Computer, South China Normal University, Guangzhou, China*
[6]*Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China*

Correspondence should be addressed to Tianyong Hao; haoty@gdufs.edu.cn and Yi Zhou; zhouyi@mail.sysu.edu.cn

Received 23 January 2018; Accepted 5 April 2018; Published 28 June 2018

Academic Editor: Javier Prieto

Natural Language Processing (NLP) empowered mobile computing is the use of NLP techniques in the context of mobile environment. Research in this field has drawn much attention given the continually increasing number of publications in the last five years. This study presents the status and development trend of the research field through an objective, systematic, and comprehensive review of relevant publications available from Web of Science. Analysis techniques including a descriptive statistics method, a geographic visualization method, a social network analysis method, a latent dirichlet allocation method, and an affinity propagation clustering method are used. We quantitatively analyze the publications in terms of statistical characteristics, geographical distribution, cooperation relationship, and topic discovery and distribution. This systematic analysis of the field illustrates the publications evolution over time and identifies current research interests and potential directions for future research. Our work can potentially assist researchers in keeping abreast of the research status. It can also help monitoring new scientific and technological development in the research field.

1. Introduction

With the development of mobile devices as well as the advances in wireless communication technologies, mobile computing is becoming a significantly important paradigm in today's world of networked computing systems [1]. Mobile computing enables a computer to be used normally while in the state of movement. Based on perceived situational information in personal and ubiquitous environments, mobile computing provides services automatically. With the rapid growth in use of mobile devices, far-reaching and diverse information is being produced rapidly and distributed instantly in digitized format [2]. A large amount of valuable information existing in unstructured texts are of great need of processing, such as web pages, short messages, Twitter/WeChat messages, etc. Natural Language Processing (NLP) focuses on the interactions between computers and natural language texts. NLP is capable of providing a computer program with the ability to process and understand unstructured texts. By automatically analyzing the meaning of user content to take appropriate actions, NLP can make applications smarter in the mobile environment.

NLP empowered mobile computing research field has attracted more and more interests from scientific community, witnessing 12 publications in 2000 to 55 publications in 2016 from Web of Science (WoS). Some representative examples are as follows. Chen et al. [3] applied the technique of multitask learning using deep neural networks to Mandarin-English code-mixing recognition. Three schemes of the auxiliary tasks were proposed to introduce the language

information to networks and to improve the prediction of language switching for the primary task of senone classification. The proposed schemes enhanced the recognition on both languages and reduced the relative overall error rates by 4.4% on average when dealing with real-world Mandarin-English corpus in mobile voice search. Ilayaraja et al. [4] presented a weighted association rule mining prefetching technique to determine the secondary service item, with the consideration of access frequency of services, semantic distance among the successive query request, and spatial distance between service instances and user context (e.g., position, service type, and query request time). Wong et al. [5] analyzed the students' vocabulary usage using a corpus analysis tool to identify and unpack the contextual conditions in which a mobile- and cloud-assisted Chinese language learning environment promoted key learning outcomes. Räsänen and Saarinen [6] proposed a method based on sparse hyperdimensional coding of sequence structures for sequence prediction. Their experiments suggested that the method was capable of capturing the relevant variable-order structure from the sequences. A NLP based tool MOTTE was developed by Puppala et al. [7] for extracting and structuring data in pathology reports automatically to support clinical solution applications. With an aim of screening information on human immunodeficiency virus/acquired immune deficiency syndrome, Adesina et al. [8] designed a monolingual short message services based system for the retrieval of frequently asked questions.

Bibliometric analysis is defined as the use of statistical methods on evaluating scholarly publications from an objective and quantitative perspective within a certain field [9]. Benefits of bibliometric analysis include (1) organizing information in a specific thematic field [10], (2) evaluating scientific developments in knowledge of a specific subject and assessing the scientific quality [11], (3) determining the impact of research funding, (4) comparing research performance across different affiliations and document changes in the research workforce, and (5) identifying emerging areas of research focus and predicting future research success [12]. As for researchers, especially newcomers, bibliometric analysis can assist them in (1) better selecting potential research topics, (2) demonstrating the values and impacts of their relevant works, (3) recognizing appropriate academic researchers to seek research collaboration, and (4) keeping abreast of new research status and new technological changes [13].

Bibliometric analysis has been widely applied to various fields for the measurement of quality and productivity of academic output and has demonstrated excellent effectiveness from long-term practice. Relevant researches mainly focused on revealing publication statistical characteristics, exploring the collaboration relationship, and uncovering research themes and their evolution. Some examples are as follows. Geng et al. [14] conducted a bibliometric survey of the research field of residential energy and greenhouse gas emissions for the purpose of uncovering research status. In their work, citation analysis was used to assess the influence of journals, countries, and authors, while network analysis was performed to evaluate the relationships among countries, authors, and keywords. Based on 117,340 obesity-related research publications indexed in Scopus database published from 1993–2012, Khan et al. [15] reported research trends and collaboration patterns in the field. Roig-Tierno et al. [16] conducted a bibliometric analysis on research publications with the application of qualitative comparative analysis (QCA). Their study revealed the differences in quantitative terms of the three variants of QCA. Albort-Morant and Ribeiro-Soriano [17] focused on the research development of business incubators. They sorted 445 publications from WoS according to bibliographic indicators such as research area and year of publication. Their study revealed the lack of publications on business incubators and highlighted the fragmented nature of research themes. Merigó and Yang [18] aimed at identifying relevant researches and the newest trends in field of operation research and management science. The analysis involved some influential journals, two hundred most cited publications, and productive and influential authors. Zhang et al. [19] quantitatively and qualitatively evaluated carbon tax related literature from 1989 to 2014 using bibliometric analysis. Their study demonstrated that the USA was the leading country and *the Vrije University Amsterdam* and *Massachusetts Institute of Technology* and *Stanford University* were the most productive affiliations in the research field. Randhawa et al. [20] conducted a systematic review of publications on open innovation (OI) research area using bibliometrics, cocitation analysis, and text mining. Three distinct areas within OI research were identified, i.e., firm-centric aspects of OI, management of OI networks, and role of users and communities in OI. In order to discover the worldwide trends in the research field of drying brick/tile, Yataganbaba and Kurtbaş [21] analyzed relevant patents in terms of, e.g., publication number, authorship and ownership, and international collaboration patterns. Merigó et al. [10] explored the research development trends in fuzzy sciences. Similar works have also been conducted in other fields, e.g., natural language processing [22], neuroimaging [23], and diabetes [24].

To the best of our knowledge, there is no scientific review of NLP empowered mobile computing research field currently. Thus, in this study, we conduct a bibliometric analysis on publications retrieved from WoS during the years 2000–2016 to explore the research status of the research field. The main objective is to address the following issues: (1) investigating publication statistical characteristics and publication collaborations, (2) exploring publication geographical distributions, (3) visualizing scientific collaboration relationships, and (4) reveling current hot research topic themes and research topic changes.

The rest of the paper is organized as follows. Section 2 introduces methods and materials. Bibliometric analysis results on retrieved research publications are reported in Section 3. Findings and discussion are shown in Section 4 while Section 5 summarizes the work.

2. Methods and Materials

Five different methods are applied to analyze research publications in the NLP empowered mobile computing field

retrieved from WoS. The details of the methods are described in Section 2.1 and the publication data is introduced in Section 2.2.

2.1. Methods

2.1.1. Descriptive Statistics Method. Descriptive statistics are brief descriptive coefficients that summarize a collection of information, which can be either a representation of the entire population or a sample. Descriptive statistics are commonly used as measures of central tendency and measures of variability. Measures of central tendency usually include mean, median, and mode, while measures of variability generally contain standard deviation, minimum and maximum variables, kurtosis, and skewness. These two measures use graphs, tables, and general discussions to simply describe data. This simplifies large amounts of data in a sensible way by presenting quantitative descriptions in a manageable form to help users understand the meaning of the data being analyzed.

In this study, descriptive statistics method was applied to acquire characteristics of the retrieved publications, including publication distribution by year, most influential publications, productive journals, authors, affiliations, and countries/regions, as well as co-authors, coaffiliation, and cocountry/region publication distribution and topic distribution by year.

2.1.2. Geographic Visualization Method. Geographic visualization or Geovisualization is a set of tools and techniques supporting the analysis of geospatial or spatial data, emphasizing knowledge construction over knowledge storage or information transmission. By combining technologies, e.g., image processing, simulation, and virtual reality, computers can help present information in a way that patterns can be found. Geovisualization can be applied to all the stages of problem-solving in geographical analysis, from development of initial hypotheses to knowledge discovery, analysis, presentation, and evaluation. According to Tobler's First Law of Geography [25], everything is related to everything else, but near things are more related than distant things. Through Geovisualization, we can use location as the key index variable and get related information which is previously unfound. Locations or extents in the earth space–time may be recorded as dates/times of occurrence. Longitude, latitude, and elevation are represented as X, Y, and Z coordinates, respectively.

In this study, we applied geographic visualization analysis to explore geographical distributions of publications in country/region level.

2.1.3. Social Network Analysis Method. Social network analysis is a process of investigating social structures using networks and graph theory [26]. It focuses on relationship structures, ranging from casual acquaintance to close bonds. Network structures are characterized in terms of nodes (items, individuals, or things within the network) with the edges or links (relationships or interactions) connecting the nodes. Researches using social network analysis have been undertaken in different areas, e.g., collaboration graphs [27], social media networks [28], and disease transmission [29]. These networks are often visualized through sociograms in which nodes are represented as points and edges are represented as lines. The social network analysis can help identify the individuals, teams, and units who play central roles, leverage peer support, and strengthen the efficiency and effectiveness of existing channels [30].

In this study, we applied social network analysis to explore the cooperation relationships for specific countries/regions, affiliations, and authors in the NLP empowered mobile computing research field. The cooperation among countries/regions, affiliations, and authors was visualized using interactive force directed networks. In the networks, nodes represented specific countries/regions, affiliations or authors, and lines indicated cooperation. The size of nodes represented publication numbers of a specific country, affiliation, or author. The width of lines reflected cooperation frequencies between two countries/regions, affiliations, or authors. The color indicated specific continent of a country/region, or specific country/region of an affiliation or author. Users could explore the cooperation relationships for specific countries/regions, affiliations, or authors by dynamically dragging the nodes.

2.1.4. Latent Dirichlet Allocation Method. Latent Dirichlet allocation (LDA), proposed by Blei [31], is a generative probabilistic model. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words, and topics are assumed to be uncorrelated.

LDA formally defines the following terms:

(1) A *word* is defined as an item from a vocabulary indexed by $\{1, \ldots, V\}$.

(2) A *document* is a sequence of N words denoted by $d = (w_1, \ldots, w_N)$.

(3) A *corpus* is a collection of M documents denoted by $D = \{d_1, \ldots, d_M\}$.

LDA assumes the following generation process:

(1) The term distribution β which contains the probability of a word occurring in a given topic is determined by $\beta \sim \text{Dirichlet}(\delta)$.

(2) The proportions θ of the topic distribution for a document d are determined by $\theta \sim \text{Dirichlet}(\alpha)$.

(3) For each word w_i in the document d, a topic is chosen by the distribution $z_i \sim \text{Multinomial}(\theta)$ and a word is chosen from a multinomial probability distribution conditioned on the topic $z_i : p(w_i \mid z_i, \beta)$.

As for variational expectation-maximization (VEM) estimation, the log-likelihood for one document $d \in D$ is given by

$$\ell(\alpha, \beta) = \log(p(d \mid \alpha, \beta))$$

$$= \log \int \left\{ \sum_{z} \left[\prod_{i=1}^{N} p\left(w_i \mid z_i, \beta\right) p\left(z_i \mid \theta\right) \right] \right\} \cdot p\left(\theta \mid \alpha\right) d\theta \tag{1}$$

Gibbs sampling defines a Markov chain in the space of possible variable assignments such that the stationary distribution of the Markov chain is the joint distribution over variables. Thus, it is a Markov Chain Monte Carlo method [32]. Its aim is to construct a Markov chain converging to the target probability distribution in the high dimensional model and then the sample distribution closest to the target probability distribution will be extracted. The log-likelihood for Gibbs sampling can be obtained through

$$\log\left(p\left(d \mid z\right)\right) = k \log\left(\frac{\Gamma\left(V\delta\right)}{\Gamma\left(\delta\right)^{V}}\right) + \sum_{K=1}^{k} \left\{ \left[\sum_{j=1}^{V} \log\left(\Gamma\left(n_K^{(j)} + \delta\right)\right) \right] - \log\left(\Gamma\left(n_K^{(\cdot)} + V\delta\right)\right) \right\}. \tag{2}$$

The perplexity, as shown in (3), is often used to evaluate the models on held-out data and is equivalent to the geometric mean per-word likelihood. The less the perplexity is, the better the model is.

$$\text{perplexity}\left(d\right) = \exp\left\{ -\frac{\log\left(p\left(d\right)\right)}{\sum_{d=1}^{D} \sum_{j=1}^{V} n^{(jd)}} \right\}. \tag{3}$$

In (4), $n^{(jd)}$ denotes how often the jth term occurs in the dth document. If the model is fitted through Gibbs sampling, the likelihood can be determined for the perplexity using

$$\log\left(p\left(d\right)\right) = \sum_{d=1}^{D} \sum_{j=1}^{V} n^{(jd)} \log\left[\sum_{K=1}^{k} \theta_K^{(d)} \beta_K^{(j)} \right] \tag{4}$$

Additionally, estimation using Gibbs sampling requires specification of values for the parameters of the prior distributions.

In this study, topic discovery and distribution were analyzed using LDA models with the following steps:

(1) We assigned the weights of segmented author keywords and Keywords Plus, publication title, and abstract as 0.4, 0.4 and 0.2, respectively, as determined in our former experiment [13].

(2) Term Frequency-Inverse Document Frequencies (TF-IDF) were used to filter out unimportant terms. As one of the most popular term-weighting schemes, TF-IDF increases proportionally to the number of times a term appears in a publication but is often offset by the frequency of the term in the whole collection of publications. We calculated the TF-IDF values of all terms to sort the terms. By manually examining these ranked terms, we defined a threshold as 0.1 empirically. Only the terms with a TF-IDF value greater than the threshold were kept for further analysis.

(3) Through sampling, 16 different topic numbers were set to $c(2:10, 15, 20, 40, 50, 80, 150, 250)$. For each topic number, 10-fold cross-validation was used to evaluate model performance. Specifically, dataset was split into 10 test datasets to conduct multiple runs. Perplexity criteria were used to select optimal topic number. α for Gibbs sampling was initialized as the mean value of α values for model fitting using VEM with the optimal topic number.

(4) With an initialized α and the optimal topic number, we adopted Gibbs sampling and VEM method to estimate the LDA model.

(5) By matching the topics detected by VEM and Gibbs sampling based on Hellinger distance, the best matches with the smallest distance could be identified. Hellinger distance is calculated as (5), in which P and Q denote two probability measures.

$$H^2\left(P, Q\right) = \frac{1}{2} \int \left(\sqrt{dP} - \sqrt{dQ}\right)^2. \tag{5}$$

2.1.5. Affinity Propagation Clustering Method. Affinity Propagation (AP) algorithm was proposed by Frey and Dueck [33]. It is a technique for data clustering based on message passing. AP does not require the predefined number of clusters. It identifies cluster centers, or exemplars as representative members of clusters. Initially, all nodes are considered as exemplars. "Preference" is used to reflect how likely one node is chosen as an exemplar. If no prior knowledge is available, all nodes will be assigned the same preference value. AP has been shown to be more efficient and effective in cluster identification than traditional clustering methods, e.g., k-means [34].

AP algorithm takes $s(i, j)$ as function of similarity to reflect the fitness of the data point j being the exemplar of data point i. The aim of AP is to maximize the similarity $s(i, j)$ between every data point i and its chosen exemplar j. Each node i also has a self-similarity $s(i, i)$. Individual data points initialized with a larger self-similarity are more likely to become exemplars. All data points are equally likely to be exemplars when they are initialized with the same constant self-similarity. The number of clusters produced will be increased and decreased accordingly with this common self-similarity input.

There are two types of messages contained in this technique. The responsibility $r(i, j)$ is directed from i to candidate exemplar j. It indicates how well suited j is to be i's exemplar, taking into consideration competing potential exemplars. The availability $a(i, j)$ is sent from candidate exemplar j back to i. It indicates j's desire to be an exemplar for i based on supporting feedback from other data points. Both the self-responsibility $r(i, i)$ and the self-availability $a(i, i)$ can

Table 1: The query used to retrieve research publications in the NLP empowered mobile computing field from WoS.

TS=((“natural language processing” OR “NLP” OR “semantic analysis” OR “bag of words” OR “word sense disambiguation” OR “named entity recognition” OR “NER” OR “sentiment analysis” OR “information extraction” “tokenization” OR “stemming” OR “lemmatization” OR “corpus” OR “stop words” OR “parts-of-speech” OR “language modeling” OR “n-grams” OR “syntactic analysis” OR “information retrieval” OR “language model”) AND (“mobile computing” OR “mobile” OR “smart device” OR “smartphone” OR “cellphone” OR “telephony device” OR “Cellular network” OR “Android” OR “iOS” OR “phone”))

reflect accumulated evidence that i is an exemplar. The update formulas for responsibility and availability are as follows:

$$r(i,j) \longleftarrow s(i,j) - \max_{j' \text{ s.t. } j' \neq j} \left\{ a\left(i,j'\right) + s\left(i,j'\right) \right\}$$

$$a(i,j) \underset{i \neq j}{\longleftarrow} \min \left\{ 0, r(i,j) + \sum_{\forall i' \notin \{i,j\}} \max \left\{ 0, r(i',j) \right\} \right\} \quad (6)$$

$$a(j,j) \longleftarrow \sum_{i' \text{ s.t. } i' \neq j} \max \left\{ 0, r\left(i',j\right) \right\}.$$

Responsibility and availability of message updates are $m_{\text{new}} = \lambda m_{\text{old}} + (1-\lambda) m_{\text{new}}$, where λ is a weighting factor between 0 and 1. In AP, the clustering is complete when the messages converge. Also, AP algorithm is able to determine when a specific data point has converged to cluster head status in its given cluster. A point becomes the cluster head when its self-responsibility plus self-availability becomes positive. Upon convergence, each node i's cluster head can be calculated using

$$CH_i = \arg\max_j \left\{ a(i,j) + r(i,j) \right\}. \quad (7)$$

In our study, with the basis of term-topic posterior probability matrix, we applied AP clustering method for the cluster analysis of the topics identified by the LDA method.

2.2. Materials. Web of Science, as the most authoritative citation database, was used as the data source for retrieving research publications in the NLP empowered mobile computing field. First of all, a list of keywords related to the “natural language processing” and “mobile computing” was determined by a domain expert. With “Science Citation Index Expanded” and “Social Sciences Citation Index” as indexes, publications used in this study were identified using the specific query in Table 1. 716 publications in “article” type during years 2000–2016 were obtained. Citations counted to September 8th, 2017 were considered for each publication.

The raw data of the 716 publications were downloaded as plain text. Key elements including title, author, journal, publication date, subject category, language, funding, author keywords, Keywords Plus, abstract, and author address, as well as number of citations, pages, and references, were extracted. In order to ensure they were closely related to the research field, manual verification was conducted by a domain expert on each publication. 471 publications were identified as relevant for analysis eventually. Further, corresponding affiliations and countries/regions were identified out from author address information. Key terms were extracted from author keywords, Keywords Plus, title, and abstract.

The statistical characteristics of the publications are shown as Table 2. The average page number of the publications is 15.66 and the average reference number of the publications is 33.29. There are 48 subject categories included, where the top 3 categories are computer science (38.76%), engineering (16.27%), and telecommunications (10.98%).

The distribution characteristics of the 471 publications are shown in Figure 1. Figure 1(a) shows the distributions of the numbers of countries/regions, affiliations, authors, and funds. Figure 1(b) shows the distributions of the numbers of keywords, pages, and references. The distribution of the number of title characters is shown in Figure 1(c). In Figure 1(d) the right bottom illustrates the distribution of the number of abstract characters.

3. Results

3.1. Publication with Year. The total publications, total citations, average number of citations per publication, and the number of annual citations are demonstrated in Figure 2. The results show that the research in the NLP empowered mobile computing field exhibits an overall upward trend in fluctuation (from 12 publications in 2000 to 55 publications in 2016). The publication number presents a stable increasing trend since 2010. Based on the data for years 2010–2016, we developed a regression model by setting the independent variables as *time/1000* and *(time/1000)*^{2}. The estimated regression model is calculated as $y = 6.7143 * 10^3 - 1.34777 * 10^4 x$. The adjusted goodness-of-fit $\overline{R^2}$ of the model is 0.9468. With the regression model, publication number in 2017 is predicted as 65, while the actual number of publications on WoS in 2017 is 66. The trend of citations does not keep step with publication number, and extreme values appear in 2002 as 431, 2007 as 503, and 2010 as 490. The average number of citations per publication is calculated as *total citations/total publications*. It shows an overall downward trend in fluctuation from 21.92 in 2000 to 2.53 in 2016. We eliminated the influence of duration since first publication using the formula: *the number of annual citations* (*C/Y*) = *total citations*/(*2016 + 1-publishing year*). The number of annual citations increases in fluctuation from 15.47 in 2000 to 139 in 2016.

3.2. Productive Journals. The top 11 contributing journals in the research field are presented in Table 3. These journals contribute about 21% of the total publications and 29.20% of the total citations. The most productive 3 are *IEEE/ACM Transactions on Audio Speech and Language Processing* (25

Table 2: The statistical characteristics of the 471 publications.

Characteristics	Statistics
Total #pub.	471
#pub. with author keywords or Keywords Plus	412
#unique publication sources	287
#unique countries (or regions)/first countries (or regions)	60; 52
#unique affiliations/first affiliations	544; 345
#unique authors/first authors/last authors	1,408; 451; 441
Average #citations	10.42
Average #countries (or regions) in one pub.	1.25
Average #affiliations in one pub.	1.64
Average #authors in one pub.	3.27
Average #funds in one pub.	0.73
Average #pages in one pub.	15.66
Average #references in one pub.	33.29
Average #author keywords or Keywords Plus	6.81
Average #words/characters in title	10.57; 80.13
Average #words/characters in abstract	186.40; 1,265.58
Language distribution	English (98.73%); Estonian (0.42%); French (0.42%); Spanish (0.21%); Afrikaans (0.21%)
Subject category distribution (Top 10)	Computer Science (38.76%); Engineering (16.27%); Telecommunications (10.98%); Acoustics (5.82%); Information Science & Library Science (2.78%); Linguistics (2.51%); Psychology (2.12%); Operations Research & Management Science (1.85%); Business & Economics (1.32%); Communication (1.32%)
Top 10 terms in author keywords and Keywords Plus	Mobile (30.36%); Information (22.08%); Retrieval (16.77%); Recognition (16.56%); System (14.86%); Speech (14.01%); Model (12.10%); Network (12.10%); Language (11.04%); Analysis (9.98%)
Top 10 terms in titles	Mobile (34.18%); Information (17.83%); System (12.53%); Retrieval (12.10%); Recognition (10.62%); Speech (8.70%); Network (8.28%); Model (7.86%); Language (7.22%); Environment (6.37%)
Top 10 terms in abstracts	Mobile (66.67%); Information (56.90%); Paper (55.41%); System (48.20%); Result (46.07%); Data (38.00%); User (38.00%); Model (37.37%); Device (32.70%); Retrieval (31.42%)

Table 3: Top 11 contributing journals in the NLP empowered mobile computing research field.

Rank	Journals	SC	TP	% *P*	TC	ACP	*H*	≥10	T100
1	IEEE/ACM Transactions on Audio Speech and Language Processing	A; E	25	5.31	447	17.88	11	12	11
2	Speech Communication	A; CS	11	2.34	179	16.27	6	6	5
3	Computer Speech and Language	CS	10	2.12	93	9.30	6	5	3
4	Expert Systems with Applications	CS; E; OR&MS	8	1.70	320	40.00	8	7	5
4	IEEE Transactions on Consumer Electronics	E; T	8	1.70	44	5.50	5	1	0
6	Mobile Information Systems	CS; T	7	1.49	95	13.57	3	2	2
6	Multimedia Tools and Applications	CS; E	7	1.49	71	10.14	3	1	1
6	Personal and Ubiquitous Computing	CS; T	7	1.49	67	9.57	4	3	1
9	Information Sciences	CS	6	1.27	85	14.17	5	3	3
10	EURASIP Journal on Wireless Communications and Networking	E; T	5	1.06	22	4.40	2	1	1
10	IEICE Transactions on Information and Systems	CS	5	1.06	11	2.20	2	0	0

Notice. Journal *IEEE Transactions on Audio Speech and Language Processing changed name as IEEE/ACM Transactions on Audio, Speech, and Language Processing* in 2013, and journal *IEEE Transactions on Speech and Audio Processing* ceased publication in 2005, and the current retitled publication is *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Therefore, publications from these 2 journals were combined as published by *IEEE/ACM Transactions on Audio, Speech, and Language Processing*; *Abbreviations*. SC: subject categories only with NLP empowered mobile computing research (A: acoustics; E: engineering; CS: computer science; OR&MS: operations research and management science; T: telecommunications); TP: total publications; % *P*: percentage of the publications; TC: total citations; ACP: average number of citations per publication, calculated as TC/TP; *H*: *H*-index; ≥10: number of publications with citations ≥10; T100: number of publications in the top 100 most influential publications.

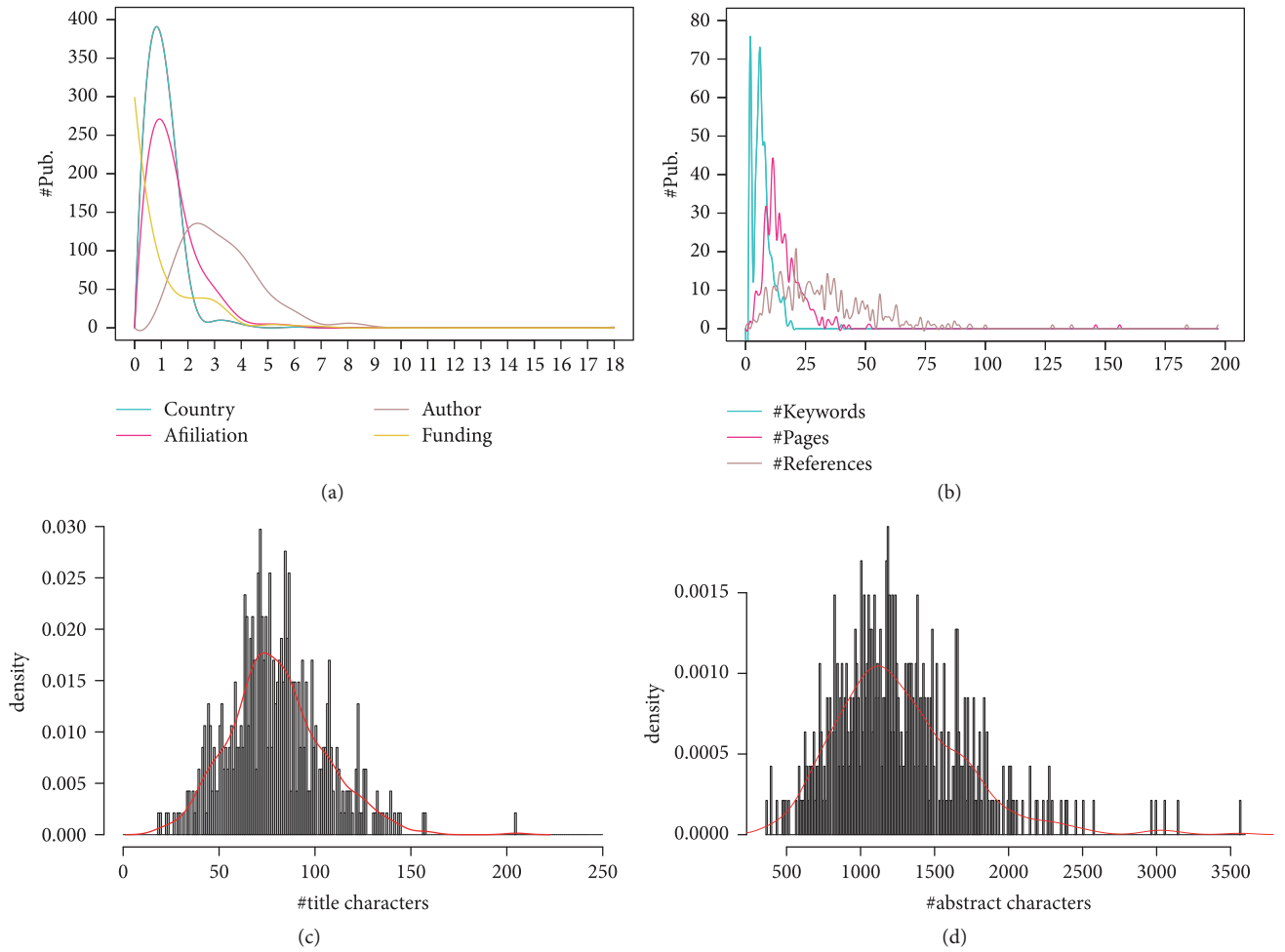


Figure 1: Distribution characteristics of the 471 publications.

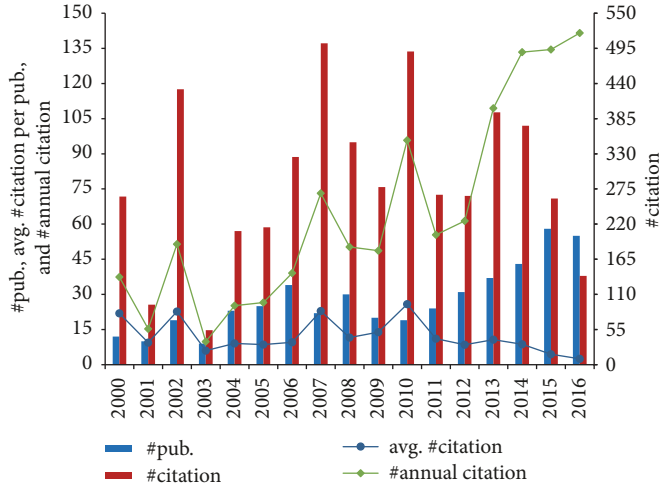
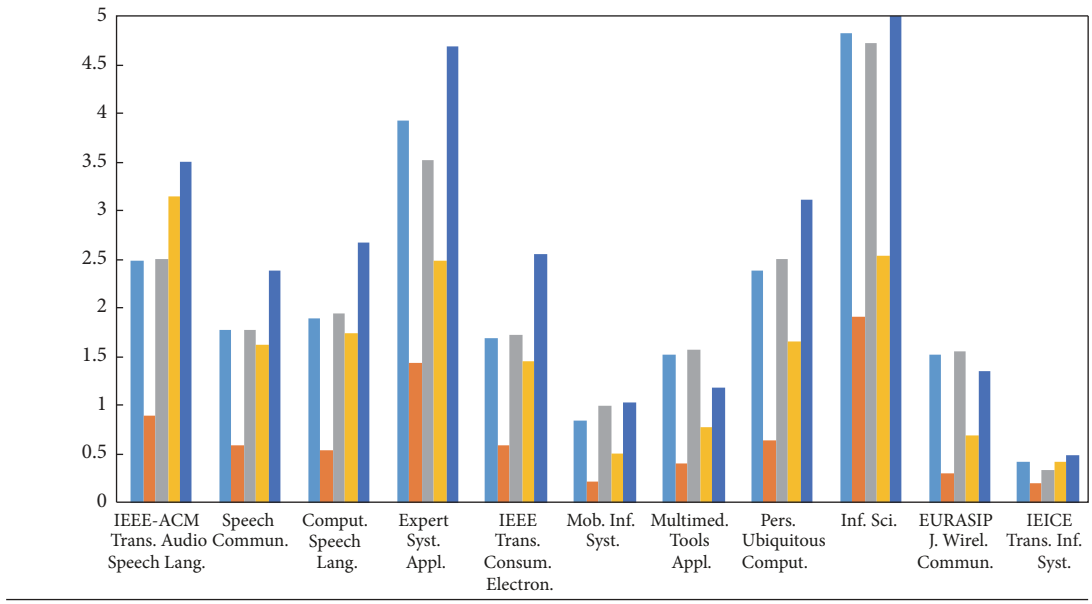


Figure 2: The statistics of the 417 publications (the light blue bars indicate total publications and the red bars indicate total citations. The dark blue line indicates average citations per publication and the green line indicates annual citations).



IF	2.491	1.768	1.9	3.928	1.694	0.849	1.53	2.395	4.832	1.529	0.411
SJR	0.813	0.593	0.536	1.433	0.587	0.219	0.396	0.648	1.91	0.304	0.191
5-Year IF	2.501	1.769	1.938	3.526	1.725	1	1.572	2.512	4.731	1.558	0.332
SNIP	3.143	1.628	1.75	2.492	1.447	0.509	0.775	1.651	2.537	0.698	0.419
CiteScore	3.5	2.38	2.67	4.7	2.56	1.03	1.19	3.12	5.37	1.36	0.49

Figure 3: Comparisons of IF, SJR, 5-Year IF, SNIP, and CiteScore for the top 11 productive journals for year 2016.

publications, 447 citations, 17.88 ACP, and 11 H-index), *Speech Communication* (11 publications, 179 citations, 16.27ACP, 6 H-index), and *Computer Speech and Language* (10 publications, 93 citations, 9.30 ACP, 6 H-index). *Expert Systems with Applications* has the highest ACP of 40.00. We found that 32 of the 100 most influential publications are published in the 11 journals. According to subject category of these 11 journals, *computer science* possesses the widest influence in the research field.

In order to better measure the overall scientific importance of these 11 journals, 5 assessment indicators acquired from Scientific Journal Rankings were used, including Impact Factor (IF), SCImago Journal Rank (SJR), 5-Year IF, Source Normalized Impact per Paper (SNIP), and CiteScore. IF is a measure for reflecting the yearly average number of citations to recent publications published in a journal. It is the primary and widely used indicator on assessing one journal's significance. SJR is a measure of scientific influence of scholarly journals. It accounts for both the number of citations received by a journal and the importance or prestige of the journals where such citations come from. 5-Year IF is calculated by dividing the number of citations to the journal in a given year by the number of publications published in that journal in the previous five years. SNIP is defined as the ratio of the journal's citation count per publication and the citation potential in its subject field. CiteScore index, launched by Elsevier in December 2016, is calculated as the ratio of total citations received in a given year by all publications published in a given journal in three previous years and the number of publications published in the journal in three previous years.

Therefore, the 11 productive journals were compared by using their IF, SJR, 5-Year IF, SNIP, and CiteScore for year 2016, as shown in Figure 3. As for IF, SJR, and CiteScore, the top 3 are *Information Sciences* (IF 4.832, SJR 1.91, and CiteScore 5.37), *Expert Systems with Applications* (IF 3.928, SJR 1.433, and CiteScore 4.7), and *IEEE/ACM Transactions on Audio Speech and Language Processing* (IF 2.491, SJR 0.813, and CiteScore 3.5). As for 5-Year IF, the top 3 are *Information Sciences* (5-Year IF 4.731), *Expert Systems with Applications* (5-Year IF 3.526), and *Personal and Ubiquitous Computing* (5-Year IF 2.512). As for SNIP score, the top 3 are *IEEE/ACM Transactions on Audio Speech and Language Processing* (SNIP 3.143), *Information Sciences* (SNIP 2.537), and *Expert Systems with Applications* (SNIP 2.492).

3.3. Most Influential Publications. The number of citations reflects the popularity and influence of a publication in the scientific community [10]. Thus, we used the total citations as a measurement of influence. There are 69 and 129 publications with the number of citations ≥20 and ≥10. Top 15 most influential publications are listed in Table 4. The publication by Miao et al. [35] in 2010 (376 citations) is the most influential one, followed by [36] published by MacKenzie and Soukoreff in 2002 (172 citations) and [37] by Strayer and Drews in 2007 (148 citations). We further consider the number of annual citations of the 15 publications. The top 3 publications measured by this indicator are [38] by Cao et al.

Table 4: Top 15 most influential publications in the NLP empowered mobile computing research field.

Rank	Title	Author/s	Year	TC	C/Y
1	Energy-Efficient Link Adaptation in Frequency-Selective Channels	Miao G. W., et al.	2010	376	53.71
2	Text Entry for Mobile Computing: Models and Methods, Theory and Practice	MacKenzie I. S.; Soukoreff R. W.	2002	172	11.47
3	Cell-Phone-Induced Driver Distraction	Strayer D. L.; Drews F. A.	2007	148	14.80
4	A Vector Space Modeling Approach to Spoken Language Identification	Li H. Z., et al.	2007	116	11.60
5	Context-Aware System for Proactive Personalized Service Based on Context History	Hong J. Y., et al.	2009	91	11.38
6	More than Words: Social Networks' Text Mining for Consumer Brand Sentiments	Mostafa M. M.	2013	88	22.00
7	The Effect of Mobility-Induced Location Errors on Geographic Routing in Mobile Ad Hoc and Sensor Networks: Analysis and Improvement Using Mobility Prediction	Son, D. J., et al.	2004	77	5.92
8	A Personalized Tourist Trip Design Algorithm for Mobile Tourist Guides	Souffriau W., et al.	2008	76	8.44
9	D'Agents: Applications and Performance of a Mobile-Agent System	Gray R. S., et al.	2002	73	4.87
10	Optical Encryption and QR Codes: Secure and Noise-Free Information Retrieval	Barrera J. F., et al.	2013	64	16.00
11	Text-Dependent Speaker Verification: Classifiers, Databases and RSR2015	Larcher A., et al.	2014	60	20.00
12	A Location-Aware Recommender System for Mobile Shopping Environments	Yang W. S., et al.	2008	59	6.56
12	An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email	Walker M. A.	2000	59	3.47
14	Landmark Recognition with Compact BoW Histogram and Ensemble ELM	Cao J. W., et al.	2016	56	56.00
14	Mobile-Agent Coordination Models for Internet Applications	Cabri G., et al.	2000	56	3.29

Abbreviations. TC: total number of citations during 2000 and 2016; C/Y: the number of annual citations.

published in 2015 ($C/Y = 56$), [35] by Miao et al. in 2010 ($C/Y = 53.71$), and [39] by Mostafa in 2013 ($C/Y = 22$). These 3 publications rank 14th, 1st, and 6th, respectively, according to total citations.

3.4. Productive Authors and Affiliations. From the 471 publications, there are 1,408 authors. 451 of them are first authors and 441 are last authors. 20 authors have 3 or more publications, and 98 authors have 2 or more publications. 20 most productive authors are listed in Table 5. According to the result, the most productive authors are *Chen, Tao* from Singapore (4 publications supported by 4 funds, 108 citations, 27 ACP, and 4 H-index) and *Mizzaro, Stefano* from Italy (4 publications, 45 citations, 11.25 ACP, and 3 H-index). *Chen, Tao* is listed as first author of 3 publications and all the 3 publications appear in top 100 most influential publications. *Mizzaro, Stefano* cooperates with others in all his 4 publications and 1 publication appears in the top 100. As for the ranking based on citation number, the top 3 productive authors are *Lee, Chin-Hui* from the USA (173 citations and 57.67 ACP), *Chen, Tao* from Singapore (108 citations and 27 ACP), and *Xie, Xing* from China (51 citations and 17 ACP). Ranking based on the ACP indicator yields the same result. *Kim, Harksoo* from South Korea achieves the most funding supports, i.e., 7 for his 3 publications.

544 affiliations from 60 countries/regions have publications in the NLP empowered mobile computing research field. Table 6 lists 15 most productive affiliations. Among them, 5 are from the USA, 3 from China, 2 from Taiwan, 1 from India, 1 from Italy, 1 from South Korea, 1 from Singapore, and 1 from England. The top 4 most productive affiliations a *Nanyang Technological University* from Singapore (8 publications, 87 citations, 10.88 ACP, and 5 H-index), *Tsinghua University* from China (8 publications, 42 citations, 5.25 ACP, and 4 H-index), *Microsoft Research Asia* from China (7 publications, 115 citations, 16.43 ACP, and 5 H-index), and *National Taiwan University* from Taiwan (7 publications, 83 citations, 11.86 ACP, and 5 H-index). *Nanyang Technological University* cooperates with others in 5 publications and serves as first affiliation in 4 of them. 3 of these 5 publications appear in the list of top 100 most influential publications. *Tsinghua University* cooperates with others in 4 publications and serves as first affiliation in all 8 publications. These 8 publications are supported by 21 funds. As for the ranking based on the total citations, the top 3 are *Georgia Institute of Technology* from the USA (550 citations and 110 ACP), *Microsoft Research Asia* from China (115 citations and 16.43 ACP), and *National Cheng Kung University* from Taiwan (62 citations and 12.4 ACP). Ranking based on the ACP indicator yields the same result.

Table 5: The most productive authors in the NLP empowered mobile computing research field.

Rank	Name	Country	TP	TC	ACP	*H*	T100	*F*	FP	LP	CP
1	*Chen, Tao*	SG	4	108	27.00	4	3	4	3	0	4
1	*Mizzaro, Stefano*	IT	4	45	11.25	3	1	0	0	0	4
2	*Baek, Jin-Wook*	KR	3	27	9.00	3	1	0	1	0	3
2	*Bertino, Elisa*	USA	3	40	13.33	3	1	3	0	2	3
2	*Cacciapuoti, Angela Sara*	IT	3	18	6.00	2	1	4	3	0	3
2	*Caleffi, Marcello*	IT	3	18	6.00	2	1	4	0	2	3
2	*Christodoulakis, Stavros*	GR	3	9	3.00	1	0	0	0	3	3
2	*Crestani, F*	UK	3	37	12.33	2	1	0	0	3	3
2	*Jung, Jason J*	KR	3	4	1.33	1	0	2	0	2	3
2	*Karanastasi, Anastasia*	GR	3	9	3.00	1	0	0	1	0	3
2	*Kazasis, Fotis G*	GR	3	9	3.00	1	0	0	0	0	3
2	*Kim, Harksoo*	KR	3	4	1.33	1	0	7	0	2	3
2	*Lee, Chin-Hui*	USA	3	173	57.67	3	3	4	1	2	2
2	*Liu, Jia*	CN	3	3	1.00	1	0	6	0	1	3
2	*Muneyasu, Mitsuji*	JP	3	4	1.33	2	0	2	1	2	3
2	*Pierre, Samuel*	CA	3	38	12.67	2	1	0	0	0	3
2	*Seide, Frank*	CN	3	48	16.00	2	2	0	0	0	2
2	*Xie, Xing*	CN	3	51	17.00	3	2	3	1	0	3
2	*Yan, Yonghong*	CN	3	9	3.00	2	0	3	0	3	3
2	*Yeom, Heon Y*	KR	3	27	9.00	3	1	0	0	3	3

Abbreviations. CA: Canada; USA: the USA; UK: England; CN: China; KR: South Korea; GR: Greece; IT: Italy; JP: Japan; SG: Singapore; TP: total publications; TC: total citations; ACP: average number of citations per publication; *H*: *H*-index; T100: number of publications in the top 100 highly cited publications; *F*: number of publications with funding; FP: number of publications as first author; LP: number of publications as last author; CP: number of collaborated publications.

Table 6: The most productive affiliations in the NLP empowered mobile computing research field.

Rank	Name	Country	TP	TC	ACP	*H*	T100	*F*	FP	CP
1	*Nanyang Technological University*	SG	8	87	10.88	5	3	1	4	5
1	*Tsinghua University*	CN	8	42	5.25	4	1	21	8	4
3	*Microsoft Research Asia*	CN	7	115	16.43	5	4	3	3	6
3	*National Taiwan University*	TW	7	83	11.86	5	3	3	5	4
5	*Georgia Institute of Technology*	USA	5	550	110.00	4	4	6	2	4
5	*Massachusetts Institute of Technology*	USA	5	10	2.00	2	0	9	4	4
5	*National Cheng Kung University*	TW	5	62	12.40	3	2	6	4	1
5	*Purdue University*	USA	5	47	9.40	4	1	4	1	5
9	*Indian Institute of Technology*	IN	4	35	8.75	3	1	3	4	1
9	*Microsoft Corporation*	USA	4	28	7.00	3	1	1	4	2
9	*The Pennsylvania State University*	USA	4	26	6.50	3	0	8	2	2
9	*Seoul National University*	KR	4	31	7.75	4	1	5	4	0
9	*University of Strathclyde*	UK	4	43	10.75	2	1	0	4	0
9	*University of Udine*	IT	4	45	11.25	3	1	0	1	3
9	*Zhejiang University*	CN	4	43	10.75	2	1	5	3	3

Abbreviations. USA: the USA; UK: England; CN: China; SG: Singapore; TW: Taiwan; IN: India; KR: South Korea; IT: Italy; TP: total publications; TC: total citations; ACP: average number of citations per publication; *H*: *H*-index; T100: number of publications in the top 100 highly cited publications; *F*: number of publications with funding; FP: number of publications as first affiliation; CP: number of collaborated publications.

3.5. Geographical Distribution. The 471 publications are from 60 countries/regions. The number of publications affiliated with 1 country/region range [61, 105), 3 countries/regions range [37, 61), and 5 range [11, 17). Table 7 shows top 15 most productive countries/regions in the field. Figure 4 illustrates geographical distributions of the publications. The top 4 countries are the USA (105 publications, 1,795 citations, 17.1 ACP, and 22 *H*-index), China (61 publications, 372 citations, 6.1 ACP, and 10 *H*-index), England (44 publications, 418 citations, 9.5 ACP, and 12 *H*-index), and South Korea (41 publications, 281 citations, 6.85 ACP, and 8 *H*-index). Among the 105 publications from the USA, 32 appear in the list of top

Table 7: The most productive countries/regions in the NLP empowered mobile computing research field.

Rank	Country	TP	TC	ACP	*H*	T100	FP (%)	Single-country/region		International collaboration	
								ACP	TP (%)	ACP	TFC (*n*)
1	USA	105	1,795	17.10	22	32	77.14	20.78	60.00	11.57	CN (12)
2	CN	61	372	6.10	10	10	91.80	4.17	57.38	9.04	USA (12)
3	UK	44	418	9.50	12	11	61.36	11.68	63.64	5.69	IE/CH (2)
4	KR	41	281	6.85	8	6	92.68	7.03	85.37	5.83	CN/USA (3)
5	TW	37	399	10.78	11	11	94.59	11.07	81.08	9.57	USA (4)
6	JP	24	77	3.21	3	1	79.17	1.44	75.00	8.50	CN (3)
7	IT	21	299	14.24	10	9	80.95	13.19	76.19	17.60	USA (3)
8	AU	18	218	12.11	7	7	61.11	18.00	38.89	8.36	USA (5)
8	CA	18	313	17.39	9	4	88.89	20.38	72.22	9.60	N/A
10	FR	17	157	9.24	6	5	64.71	4.45	64.71	18.00	CN/USA (2)
10	GR	17	38	2.24	3	0	100.00	2.24	100.00	0.00	N/A
10	ES	17	124	7.29	7	2	88.24	6.43	82.35	11.33	USA (2)
13	SG	16	355	22.19	9	7	75.00	14.90	62.50	34.33	USA (2)
14	HK SAR	15	98	6.53	6	2	53.33	9.17	40.00	4.78	CN/USA (4)
15	DE	14	114	8.14	5	3	85.71	8.11	64.29	8.20	CN (12)

Abbreviations. USA: America; UK: England; CN: China; SG: Singapore; TW: Taiwan; KR: South Korea; IT: Italy; JP: Japan; AU: Australia; CA: Canada; FR: France; GR: Greece; ES: Spain; HK SAR: Hong Kong SAR; DE: Germany; IE: Ireland; CH: Switzerland; TP: total publications; TC: total citations; ACP: average number of citations per publication; *H*: *H*-index; T100: number of publications in the top 100 highly cited publications; FP (%): percentage of publications as first affiliation; TFC (*n*): number of cooperation times with the closest collaborator, where $n \geq 2$.

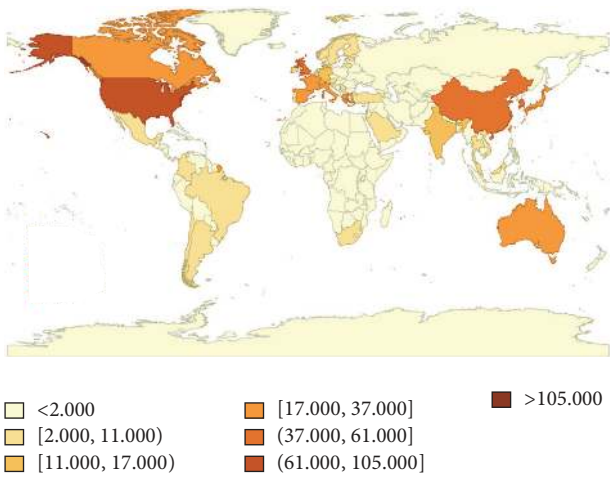


Figure 4: Geographical distributions of the NLP empowered mobile computing research publications.

100 most influential publications. It is noted that publications from Singapore have the highest ACP, which indicates the high quality of the publications. As for most of the top 15 productive countries/regions, the international collaboration rates are around 30%, except for Greece with 0 and Australia with 61.11%. The USA is the closest collaborator for 9 of the 15 countries/regions. The ACP of internationally collaborated publications is much higher than that of noninternationally collaborated publications for countries/regions like China, Japan, Italy, France, Spain, and Singapore. This potentially indicates that international collaboration can improve the quality of their publications.

Since the publications are mainly distributed in the USA, China, England, and South Korea, we further explored the annual publication distributions for these 4 countries, as shown in Figure 5. The number of publications for the USA and China is on the whole presenting upward trend in fluctuation. As for the USA, the number increases from 2 in 2000 to 9 in 2007 but dwindles to 2 in 2010. After that, the upward trend becomes more significant. The situation for China is quite like that for the USA after 2010, witnessing the great mass upsurge on the NLP empowered mobile computing research in these two countries since 2010. As for England and South Korea, the number of publications does not increase much in fluctuation with years going on.

3.6. Cooperation Relationship. Figure 6 shows the trends of the international collaborative and the percentage of international collaborative publications. We found that the international collaborative publications increase during the years 2000–2016. The percentage of international collaborations increases from 8.33% in 2000 to 32.73% in 2016. This indicates that international collaborations in the NLP empowered mobile computing research field have become increasingly important.

Figures 7 and 8 present the institutional level of cooperation and the author level of cooperation, respectively. The cooperation between different institutions is becoming more and more frequent. The percentage of institution-collaborative publication increases from 16.67% in 2000 to 58.18% in 2016. More than 90% of the publications are multiauthored since 2011. It is worth noticing that the percentage reaches up to 100% in 2015.

Furthermore, the cooperation relations for specific countries/regions, affiliations, and authors were visualized with

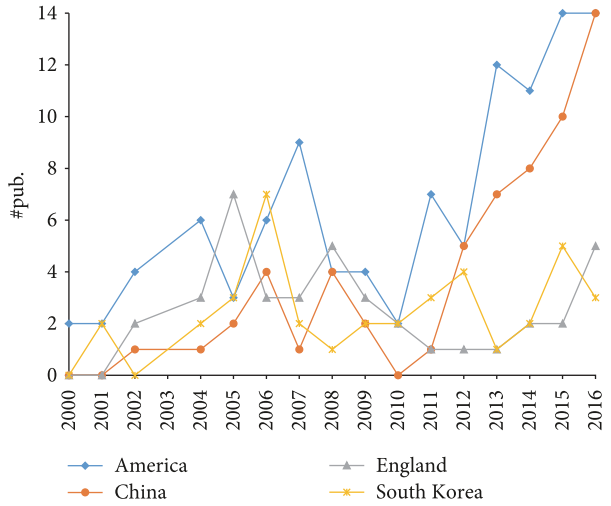


Figure 5: Publication distributions by year for the top 4 countries/regions.

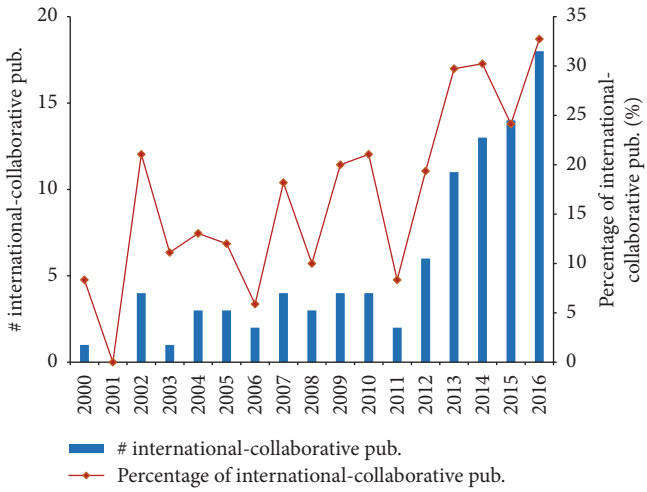


Figure 6: International collaborative publication distribution by year.

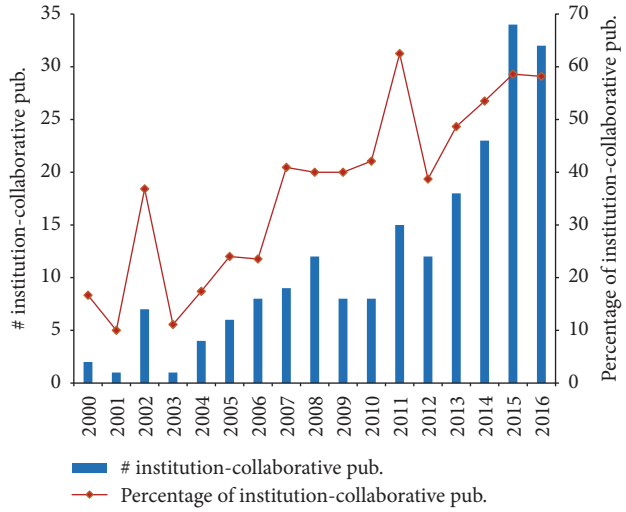


Figure 7: Institution-collaborative publication distribution by year.

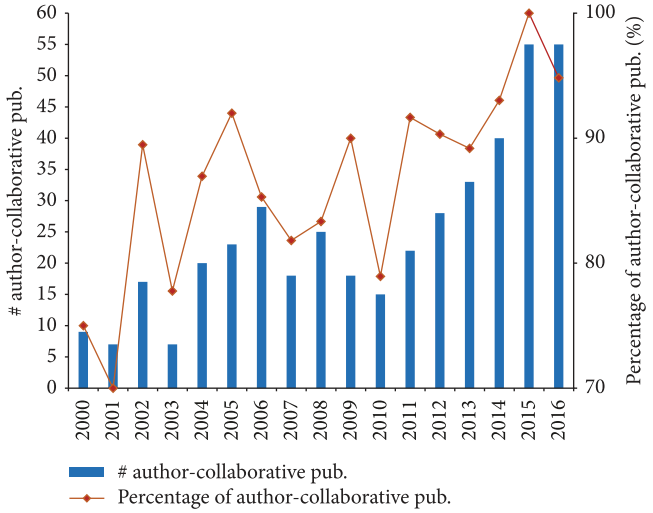


Figure 8: Author-collaborative publication distribution by year.

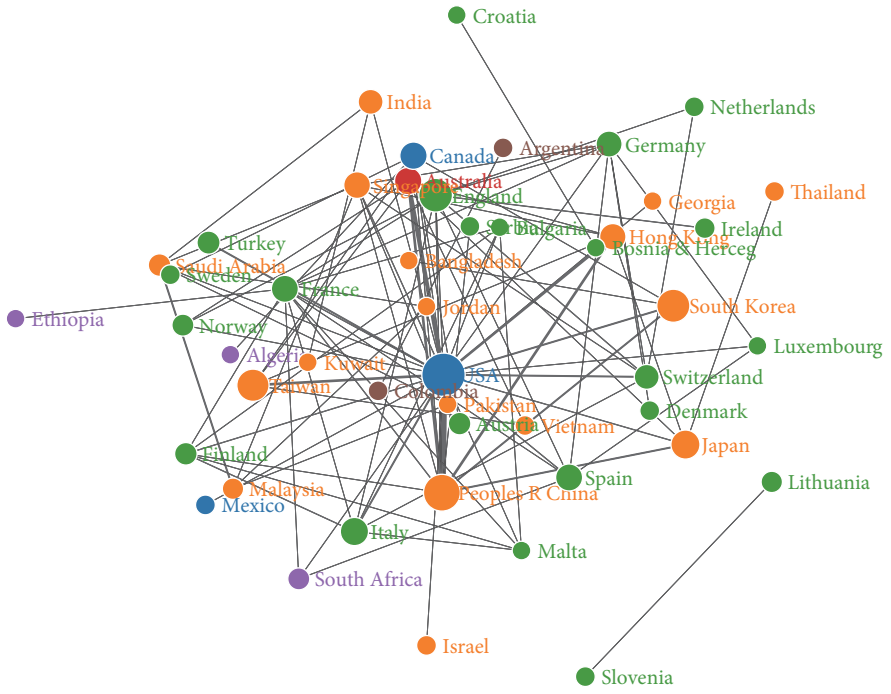


Figure 9: Cooperation network of 48 countries/regions (node colors represent different continents, e.g., orange for Asia, blue for North America, green for Europe, red for Oceania, purple for Africa, and brown for South America). The network can be accessed via the link (http://www.zhukun.org/haoty/resources.asp?id=NLPEMC_cocountry).

social network analysis. A cooperation network for 48 countries/regions is shown in Figure 9. 17 of them come from Asia (represented as orange nodes), 3 from North America (represented as blue nodes), 22 from Europe (represented as green nodes), 3 from Africa (represented as purple nodes), 2 from South America (represented as brown nodes), and 1 from Oceania (represented as red node). There are 141 affiliations with the number of publications ≥ 2, and there exists cooperation among 91 of them. Figure 10 shows a cooperation network of the 91 affiliations. 23 of the 91 affiliations are from the USA and 14 from China. As for cooperation of author level, there are 98 authors with publication count ≥ 2. among them, 65 authors involve in cooperation. We created a cooperation network of the 65 authors, as shown in Figure 11.

3.7. Topic Discovery and Distribution. By setting TF-IDF value threshold as 0.1, the terms were ranked by frequency. Table 8 lists top 20 most frequent terms, in which the top 5

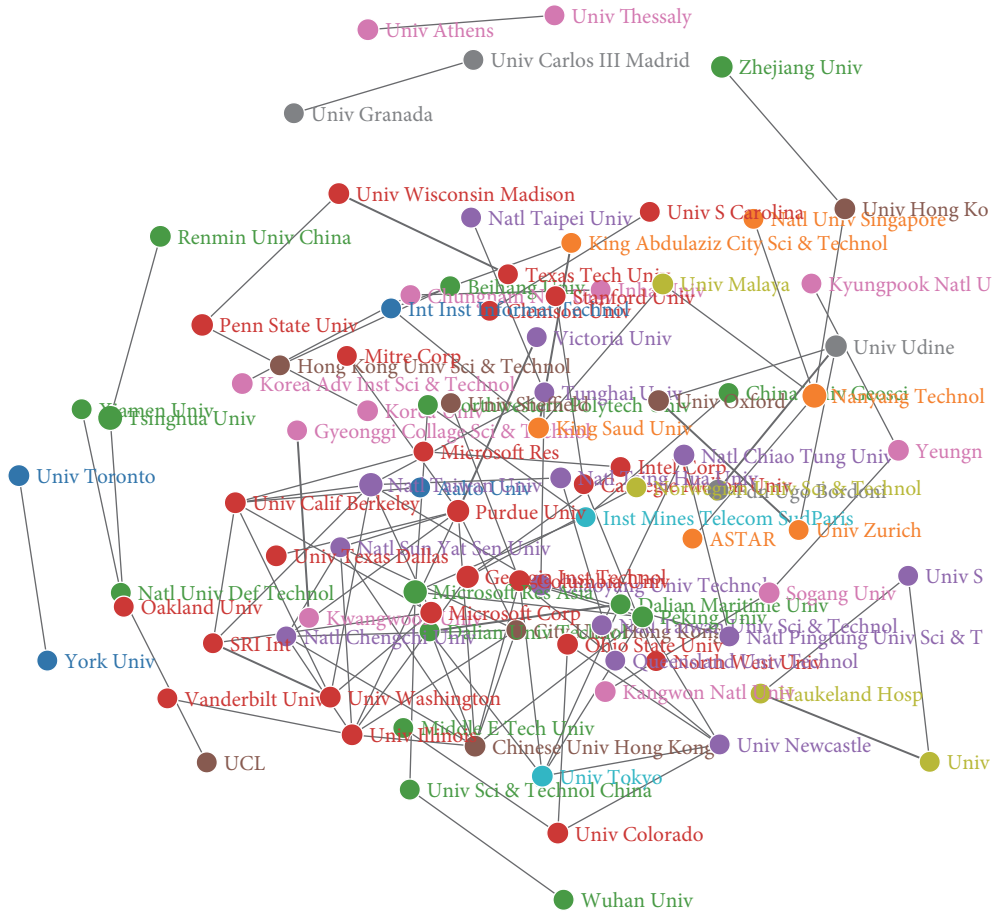


Figure 10: Cooperation network of 91 affiliations (node colors represent different countries/regions, e.g., red for the USA, pink for South Korea, and purple for Australia). The network can be accessed via the link (http://www.zhukun.org/haoty/resources.asp?id=NLPEMC_coaffiliation).

terms are "Agent" (369), "Image" (215), "Sentiment" (128), "Dialogue" (83), and "Health" (81). Figure 12 presents the perplexities of models fitted by using Gibbs sampling with different numbers of topics. The result suggests that the optimal topic number is between 40 and 80. Hence, we set the topic number as 40. The α was set to the mean value 0.01101332 in the cross-validation fitted using VEM. Using the parameters, we estimated the LDA model using Gibbs sampling. By semantics analysis of representative terms in each topic, as well as reviewing text intention of the corresponding publications, we assigned potential theme to each topic. The order of topics are determined based on Hellinger distance. Specifically, Topic 36 is the best matching topic and Topic 11 ranks 2nd, while Topic 37 is the worse matching one. Due to space limitation, Table 9 only displays the top 10 best matching topics with the most frequent terms. Each publication was assigned to the most likely topic with the highest posterior probability. Integrating topic proportions for all the publications, we obtained a topic distribution. The 4 most frequent research topics are Topic 36 (6.38%), Topic 4 (4.26%), Topic 11 (3.83%), and Topic 17 (3.83%), while the 4 least frequent research topics are Topic 26 (1.49%), Topic 23 (1.28%), Topic 10 (1.06%), and Topic 20 (1.06%).

We used the AP clustering analysis to perform the cluster analysis of the 40 topics. One way for measuring topic similarity is based on term-level similarity with the hypothesis that topics may contain the same terms. The clustering result based on term-topic posterior probability matrix is shown in Figure 13, where the 40 topics are categorized into 8 groups.

Identifying emerging research topics can provide valuable insights into the development of the research field. Likewise, identification of fading research topics can also help understand the hot spots evolution [40]. We then explored the annual publication proportions of the 40 research topics, as shown in Figure 14. We used Mann–Kendall test [41], a nonparametric trend test, to examine whether increasing or decreasing trends are existing in the 40 topics. Test results show that 12 topics, including Topic 1, Topic 4, Topic 7, Topic 10, Topic 14, Topic 18, Topic 20, Topic 26, Topic 29, Topic 32, Topic 33, and Topic 39, present a statistically significant increasing trend. While Topic 36 presents a statistically significant decreasing trend, both at the two-sided $p = 0.05$ levels.

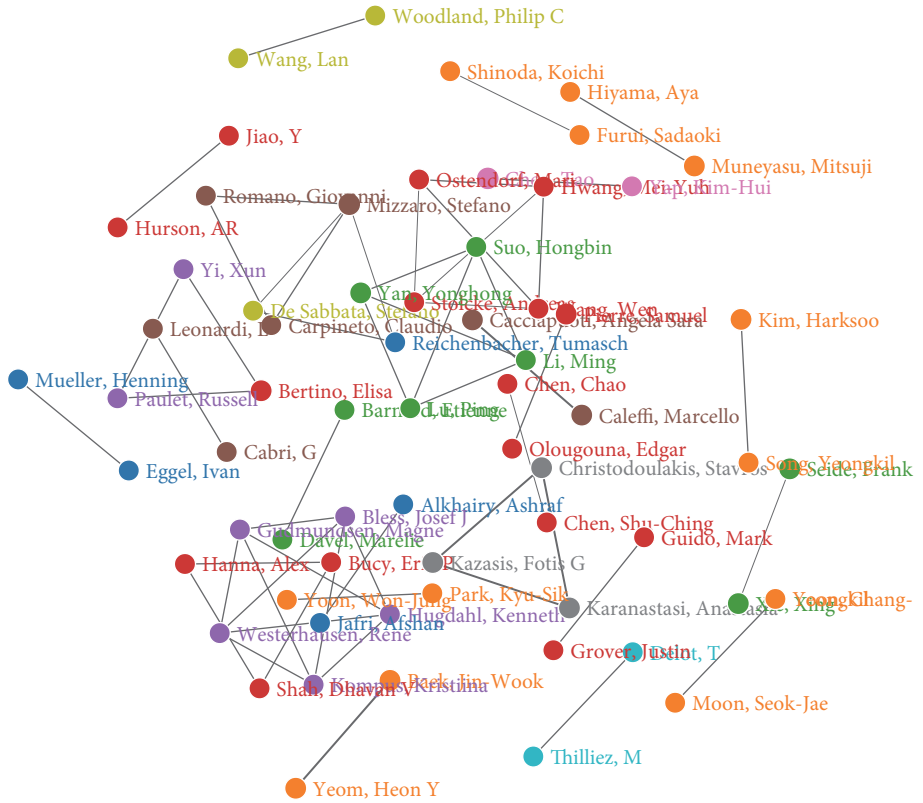


Figure 11: Cooperation network of 65 authors (node colors represent different countries/regions, e.g., range for South Korea, red for the USA, purple for Australia, green for China, and brown for Italy). The network can be accessed via the link (http://www.zhukun.org/haoty/resources.asp?id=NLPEMC_coauthor).

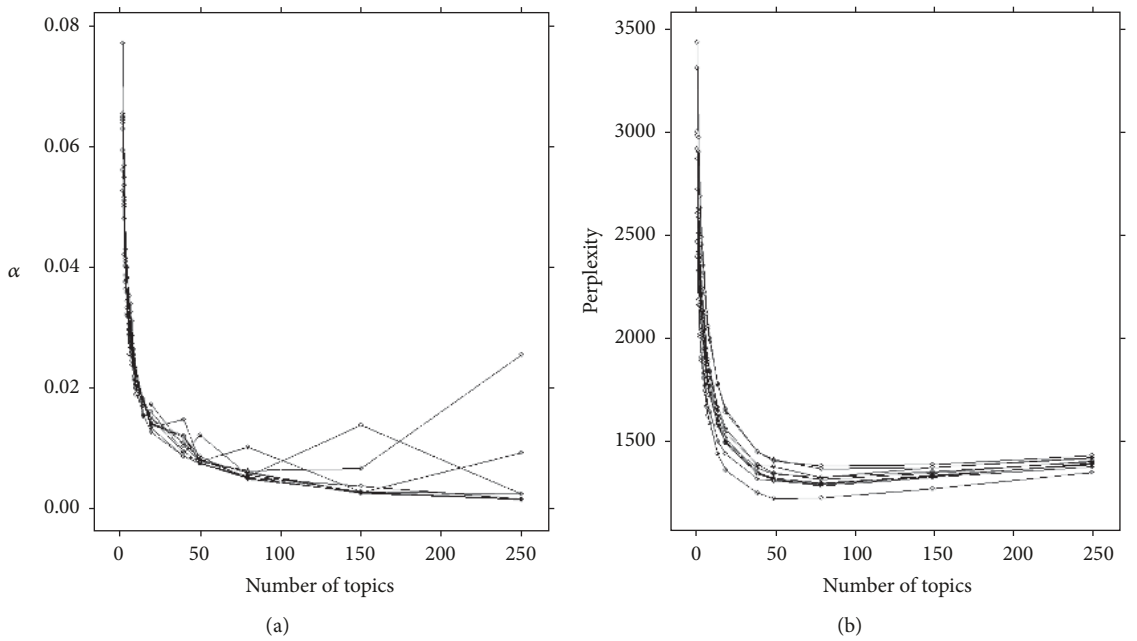


Figure 12: (a) Estimated α value for the models fitted using VEM. (b) Perplexities of the test data for the models fitted by using Gibbs sampling. Each line corresponded to one of the folds in the 10-fold cross-validation.

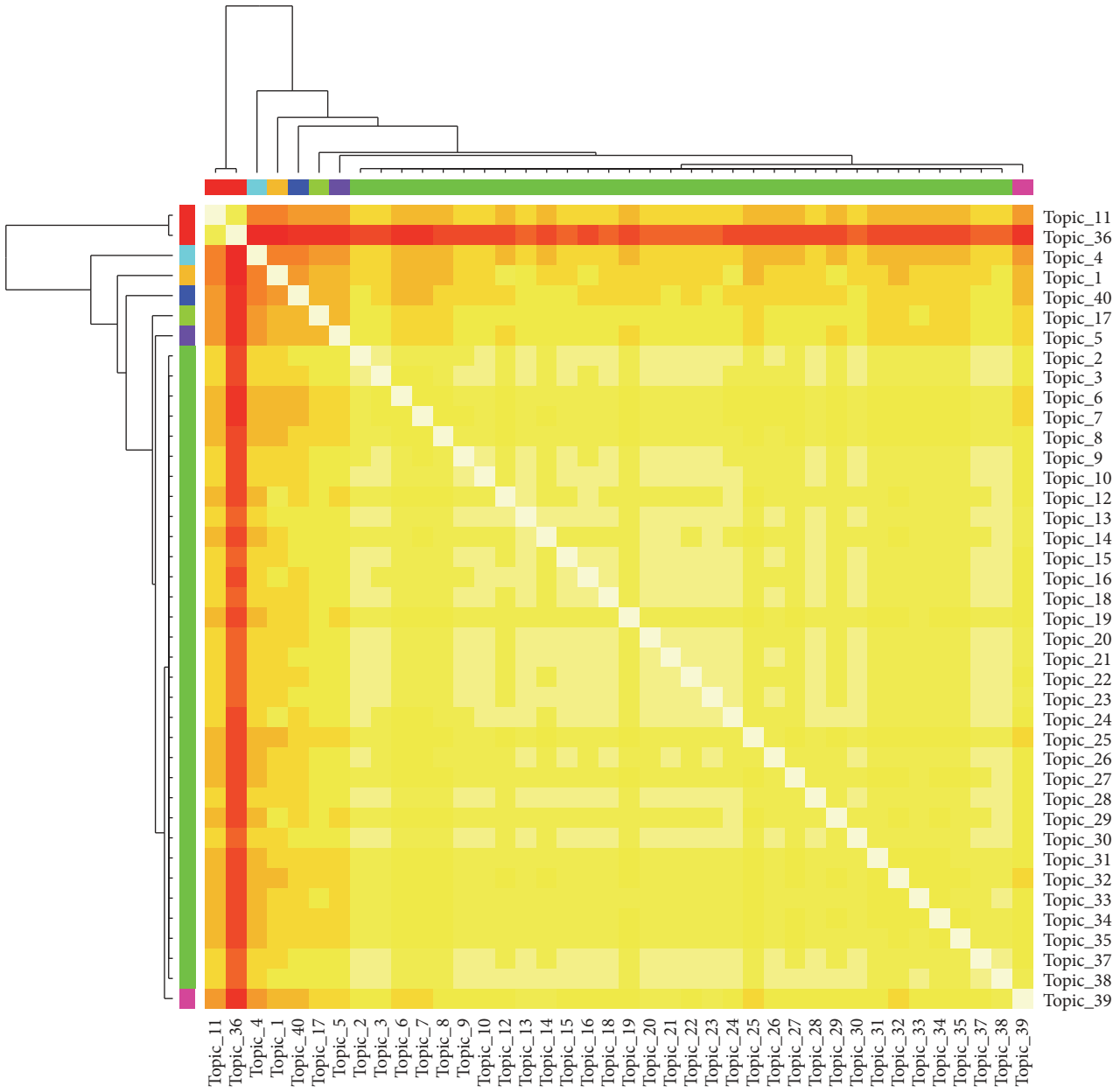


FIGURE 13: The visualized result of hierarchical clustering based on term-topic posterior probability matrix.

4. Discussions

This study provides a most up-to-date bibliometric analysis on the publications in WoS during the years 2000–2016 in the NLP empowered mobile computing research field. Some interesting findings are discussed below.

The annual number of the publication distribution shows a significant growth trend, from 12 publications in 2000 to 55 publications in 2016. This indicates a growing interest in the research field.

The literature characteristics analysis shows that the 417 publications are widely dispersed throughout 287 journals. 11 most productive journals together contribute about 21% of the total publications. The top 3 are *IEEE/ACM Transactions on Audio Speech and Language Processing*, *Speech Communication*, and *Computer Speech and Language*. *Computer science* is the most shared subject among these 11 journals. Journal *Information Sciences* possesses the highest IF, SJR, 5-Year IF, and CiteScore, except for the SNIP score in year 2016.

Top 3 most influential publications are: [35] by Miao et al. published in 2010, [36] by MacKenzie and Soukoreff published in 2002, and [37] by Strayer and Drews published in 2007.

There are 1,408 authors and 544 affiliations involved in the publications. Most authors (79.18%) have only 1 publication, and 4.25% of the authors have 3 or more publications. The most productive authors are *Chen, Tao* from Singapore and *Mizzaro, Stefano* from Italy. In addition, most affiliations

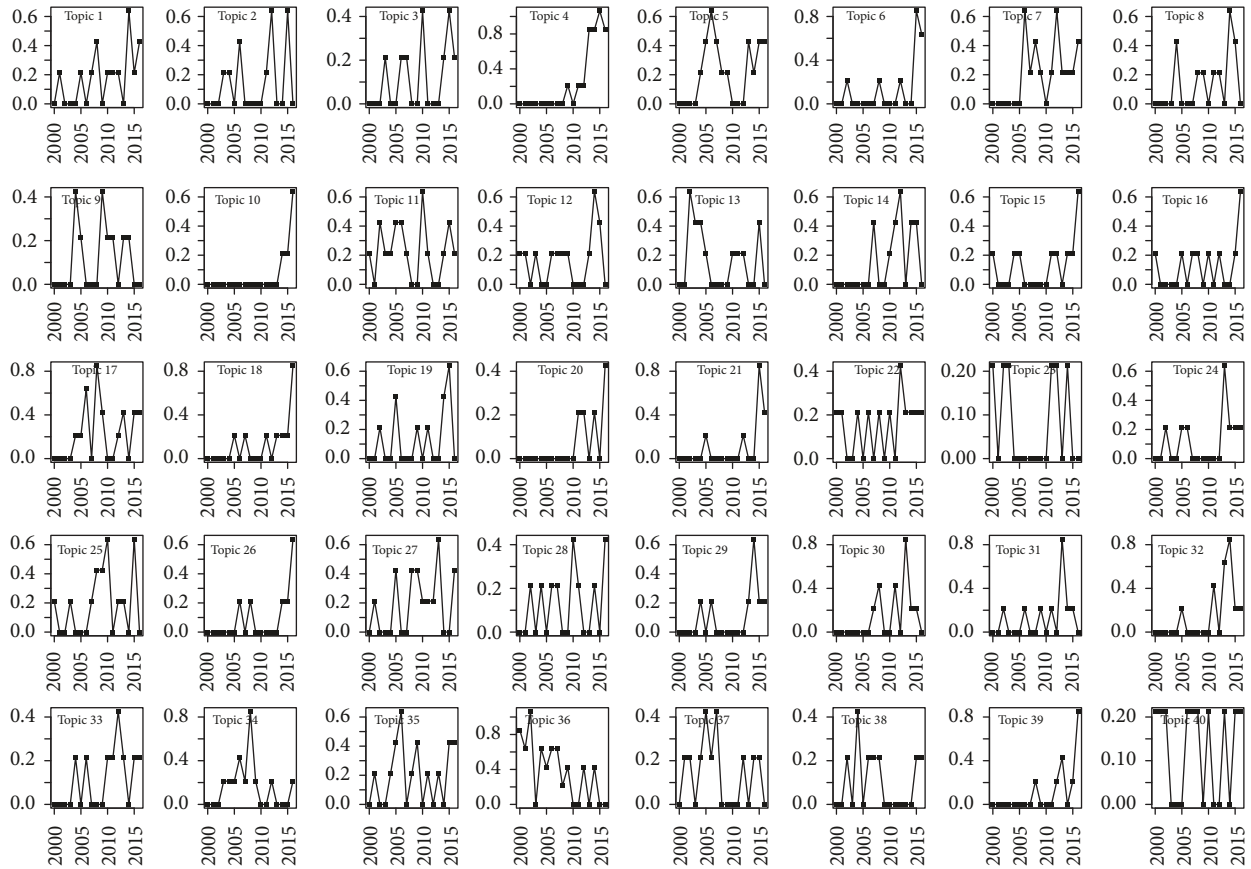


Figure 14: The trends of the 40 research topics during 2000–2016 (x-coordinate as year, y-coordinate as proportion %).

Table 8: Top 20 most frequent terms.

Rank	Stemmed terms	Occurrence number		
		Total	2000–2008	2009–2016
1	Agent	369	250	119
2	Image	215	70	145
3	Sentiment	128	0	128
4	Dialogue	83	49	34
5	Health	81	2	79
6	Music	76	27	49
7	Radio	74	10	64
8	Unit	74	51	23
9	Adaptation	70	40	30
10	Relevance	69	29	40
11	Geographic	66	37	29
12	Short Messages	66	9	57
13	Protocol	65	20	45
14	Chinese	64	29	35
15	Medical	60	16	44
16	Recommendation	60	4	56
17	Clustering	54	20	34
18	Privacy	54	9	45
19	Ad hoc	53	9	44
20	Traffic	52	17	35

(70.06%) have 1 publication. 11.89% of the affiliations have 3 or more publications. The most productive affiliations are *Nanyang Technological University* from Singapore and *Tsinghua University* from China. *Lee, Chin-Hui* from USA with 57.67 ACP ranks 1st among top 20 productive authors, and *Georgia Institute of Technology* from USA with 110 ACP ranks 1st among 15 most productive affiliations.

Through geographic visualization analysis, 60 countries/regions have participated in the publications. The top 15 productive countries/regions are developed countries/regions, except for China. As the top 2, the USA and China have shown a significant growth in the numbers of scientific publications since 2010. These numbers are predicted to continue to increase in the coming years. This partially reflects the need of the development of NLP techniques in solving mobile computing issues.

Scientific collaboration analysis shows that there are significant growth of international collaborations, institution-collaborations as well as author-collaborations. Through social network analysis, we found that researchers tend to collaborate with others within the same country or area, with institutions under similar administration, or with a neighboring country or area. However, some research institutions might have separate administration arrangements from their associated universities or hospitals and a researcher might be affiliated with multiple institutions. The co-authors

Table 9: Top 15 most frequent terms for the top 10 best matching topics.

Topic	Potential theme	Top high frequency terms
36	Mobile agent computing	Agent; Coordination; Java; Migration; Protocol; Mobile-agent; Failure; Itinerary; Filtering; Turkish; Attack; Commerce; Context-aware; Truncation; Crash
11	Mobile agent computing	Agent; Planning; Ontology; Cloud; Multi-agent; Net; Interoperability; Neural; Peer-to-Peer; Broadband; Instruction; Complementarity; Natural Language; Traffic; Grounding
32	Mobile privacy and security	Privacy; Private; Secure; Location-Based Services; Encryption; Points of Interest; Protection; Approximate; Attack; Path; Privacy-preserving; Streaming; Password; Protocol; Cryptosystem
1	Image and syllable events	Image; Particular Allophones; Re-ranking; Composite Phoneme; Simple Phonemes; Syllable; Thing; iPad; On-Premise Signs; Spreading; Bow; Modern Orthography; Arabic; Content-based; Descriptor
4	Mobile social media computing	Sentiment; Opinion; Twitter; Tweet; Customer; Suggestion; Emojis; Emotion; Micro-blog; Protest; Brand; Suggestive; Microblog; Orientation; Box
8	Mobile radio	Radio; Phone-in; Localization; Australian; Formulation; Island; Reporting; Talkback; Involvement; Caller; Dialogic; Stance; Backlinking; Cloud; French
5	Mobile location computing	Geographic; Relevance; Seeking; Innovation; Subspace; Tourism; Birthright; Firm; Flier; Sensing; TILES (Temporal, Identity, Location, Environmental and Social); Cross-space; Location-aware; Personalized; Reposting
40	Context-aware computing	Dialogue; Context-aware; Estonian; Clarification; Array; Problematic; Reformulation; Verbose; Email; Mobile Information Services enabled by Mobile Publishing; Non-understanding; Publishing; Agent; Directive; Reinforcement
10	Second screen response	Gesture; Debate; PreFrontal Cortex; Adult; Presidential; Walking; Facial; Twitter; Educational; Gait; Political; Touch; Biometrics; Blink; Cortex
35	Language learning and modeling	Chinese; Information Retrieval; Peer-to-Peer; Conditional Random Field; Update; Apprentice; Affordances; Disyllabic; Website; Workplace; Self-study; Skip-chain; Descriptive; Mobile Peer-to-Peer; Multilingual

might actually work together but are affiliated with different institutions. Therefore, it is worth noticing that institution-wise collaboration might not be the actual collaboration among institutions.

Most topics identified using LDA method are recognizable, as they are related to major issues in the research field. Due to space constraints, here we only provide interpretations of some representative topics.

Topic 36 and Topic 11 contain words such as "Agent", "Mobile-agent", "Multi-agent", "Itinerary", "Migration", "Protocol", and "Truncation". Thus, Topic 36 and Topic 11 pertain to *mobile agent computing*. As an emerging and exciting paradigm for mobile computing applications [42], mobile agent can not only support mobile computers and disconnected operations but also provide an efficient, convenient and robust programming paradigm for implementing distributed applications. The use of mobile agent can bring about significant benefits, e.g., reduction of network traffic, overcoming network latency, and seamless system integration. Therefore, mobile agent is well adapted to the domain of mobile computing.

Topic 32 discusses *events about mobile privacy and security*. Words in this topic include "Privacy", "Private", "Secure", "Encryption", "Privacy-preserving", "Password", and "Cryptosystem". As pointed out by Mollah et al. [43], security and privacy challenges are introduced with the development of mobile cloud computing which aims at relieving challenges of the resource constrained mobile devices in mobile computing area. Studies centering on mobile privacy can be found. For example, Xi et al. [44] applied Private Information Retrieval techniques in finding the shortest path between an origin and a destination in location privacy issues without the risk of disclosing their privacy.

Topic 1 discusses *mobile computing on image and syllable events*. It includes words such as "Image", "Syllable", "Re-ranking", "Content-based", "Composite Phoneme", "Simple Phonemes", and "Modern Orthography". Image search in mobile device is quite worthy of challenge [45]. Many researchers are seeking ways to solve this problem. For example, Cai et al. [46] presented a new geometric reranking algorithm specific for small vocabulary in aforementioned scenarios based on Bag-of-Words model for image retrieval. Mobile computing on syllable events is another focus. A representative work is by Eddington and Elzinga [47]. They conducted a quantitative analysis on the phonetic context of word-internal flapping with great attention paid to stress placement, following phone, and syllabification.

Topic 4 mainly focuses on *mobile social media event*. Words like "Twitter", "Sentiment", "Tweet", "Emojis", "Micro-blog", "Opinion", "Public", and "Emotion" can be found within this topic. With the rapid development of social network, information spreading and evolution is facilitated with popularity of the environment of wireless communication, especially social media platform on mobile terminals [48]. Researchers are gradually paying attention to this area. For example, based on 100 million collected

messages from Twitter, Wang et al. [49] presented a hybrid model for sentimental entity.

Based on topic distributions, we found that *mobile agent computing*, *mobile social media computing*, and *sound related event computing* are 3 highest-frequent research themes. From Figure 14 as well as Mann–Kendall test results, we found that some research themes present a statistically significant increasing trend, e.g., *image and syllable related events*, *mobile social media computing*, and *healthy related events*, while researches on *mobile agent computing* presents a statistically significant decreasing trend.

In the thematic analysis, the optimal number of topics was selected as 40 by a statistical measure of model fitting the data. However, mechanical reliance on statistical measures might lead to the selection of a less meaningful topic model [50]. Hence, we manually checked the robustness of the results by confirming identified topics using a qualitative assessment with the basis of prior knowledge. For each topic, we checked the semantic coherence of its high-frequency terms and examined the contents of publication with a high proportion of this topic.

Through the AP clustering analysis on the 40-topics, 8 clusters were identified, i.e., *mobile agent computing, mobile social media computing, image and syllable related events, context-aware computing, sound related events, mobile location computing, healthy related events*, and other events. The results of AP clustering analysis are on the whole sensible and easy-to-understand. However, we still found that the 8 categories vary a lot in topic numbers. One possible reason is the choice of clustering method. We then adopted hierarchical clustering method with category number setting to 8. The result was similar with AP clustering. Another possible reason is the sample size since the number of the relevant publications in WoS is limited.

This study is the first to thoroughly explore research status of the NLP empowered mobile computing research field in the statistical perspective. The study provides a comprehensive overview and an intellectual structure of the field from 2000 to 2016. The findings can potentially help researchers especially newcomers systematically understand the development of the field, learn the most influential journals, recognize potentially academic collaborators, and trace research hotspots.

For future work, there are several directions. First, more comprehensive data is expected to be included. Though WoS is a widely applied repository for bibliometric analysis due to its high authority, some relevant conference proceedings have not been indexed yet in WoS. Second, we intend to employ different data clustering methods and compare clustering results for deeper cluster analyzing.

5. Conclusions

We conducted a bibliometric analysis on natural language processing empowered mobile computing research publications from Web of Science published during years 2000–2016. The literature characteristics were uncovered using a descriptive statistics method. Geographical publication distribution was explored using a geographic visualization method. By applying a social network analysis method, cooperation relationships among countries/regions, affiliations, and authors were displayed. Finally, topic discovery and distribution were presented using a LDA method and an AP clustering method. We believe the analysis can help researchers comprehend the collaboration patterns and distribution of scholarly resources and research hot spots in the research field more systematically.

Disclosure

Tianyong Hao and Yi Zhou are the corresponding authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work was substantially supported by the grant from National Natural Science Foundation of China (no. 61772146), the Innovative School Project in Higher Education of Guangdong Province (No. YQ2015062), Science and Technology Program of Guangzhou (no. 201604016136), and Major Project of Frontier and Key Technical Innovation of Guangdong Province (no. 2014B010118003).

References

[1] G. Deepak and B. S. Pradeep, "Challenging issues and limitations of mobile computing," vol. 3, pp. 177–181, 2012.

[2] K.-Y. Chung, J. Yoo, and K. J. Kim, "Recent trends on mobile computing and future networks," *Personal and Ubiquitous Computing*, vol. 18, no. 3, pp. 489–491, 2014.

[3] M. Chen, J. Pan, Q. Zhao, and Y. Yan, "Multi-task learning in deep neural networks for Mandarin-english code-mixing speech recognition," *IEICE Transaction on Information and Systems*, vol. E99D, no. 10, pp. 2554–2557, 2016.

[4] N. Ilayaraja, F. Mary Magdalene Jane, M. Safar, and R. Nadarajan, "WARM Based Data Pre-fetching and Cache Replacement Strategies for Location Dependent Information System in Wireless Environment," *Wireless Personal Communications*, vol. 90, no. 4, pp. 1811–1842, 2016.

[5] L.-H. Wong, R. B. King, C. S. Chai, and M. Liu, "Seamlessly learning Chinese: contextual meaning making and vocabulary growth in a seamless Chinese as a second language learning environment," *Instructional Science*, vol. 44, no. 5, pp. 399–422, 2016.

[6] O. J. Räsänen and J. P. Saarinen, "Sequence prediction with sparse distributed hyperdimensional coding applied to the analysis of mobile phone use patterns," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 9, pp. 1878–1889, 2016.

[7] M. Puppala, T. He, S. Chen et al., "METEOR: An Enterprise Health Informatics Environment to Support Evidence-Based Medicine," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2776–2786, 2015.

[8] A. O. Adesina, K. K. Agbele, A. P. Abidoye, and H. O. Nyongesa, "Text messaging and retrieval techniques for a mobile health

information system," *Journal of Information Science*, vol. 40, no. 6, pp. 736–748, 2014.

[9] W.-T. Chiu and Y.-S. Ho, "Bibliometric analysis of tsunami research," *Scientometrics*, vol. 73, no. 1, pp. 3–17, 2007.

[10] J. M. Merigó, A. M. Gil-Lafuente, and R. R. Yager, "An overview of fuzzy research with bibliometric indicators," *Applied Soft Computing*, vol. 27, pp. 420–433, 2015.

[11] D. Bouyssou and T. Marchant, "Ranking scientists and departments in a consistent manner," *Journal of the Association for Information Science and Technology*, vol. 62, no. 9, pp. 1761–1769, 2011.

[12] A. Mazloumian, "Predicting Scholars' Scientific Impact," *PLoS ONE*, vol. 7, no. 11, Article ID e49246, 2012.

[13] X. Chen, H. Xie, F. Wang, Z. Liu, J. Xu, and T. Hao, "Natural Language Processing in Medical Research: A Bibliometric Analysis," *BMC Medical Informatics and Decision Making*, vol. 18, supplement 1, no. 14, 2018.

[14] Y. Geng, W. Chen, Z. Liu et al., "A bibliometric review: Energy consumption and greenhouse gas emissions in the residential sector," *Journal of Cleaner Production*, vol. 159, pp. 301–316, 2017.

[15] A. Khan, N. Choudhury, S. Uddin, L. Hossain, and L. A. Baur, "Longitudinal trends in global obesity research and collaboration: A review using bibliometric metadata," *Obesity Reviews*, vol. 17, no. 4, pp. 377–385, 2016.

[16] N. Roig-Tierno, T. F. Gonzalez-Cruz, and J. Llopis-Martinez, "An overview of qualitative comparative analysis: A bibliometric analysis," *Journal of Innovation Knowledge*, vol. 2, no. 1, pp. 15–23, 2017.

[17] G. Albort-Morant and D. Ribeiro-Soriano, "A bibliometric analysis of international impact of business incubators," *Journal of Business Research*, vol. 69, no. 5, pp. 1775–1779, 2016.

[18] J. M. Merigó and J.-B. Yang, "A bibliometric analysis of operations research and management science," *OMEGA - The International Journal of Management Science*, vol. 73, pp. 37–48, 2017.

[19] K. Zhang, Q. Wang, Q.-M. Liang, and H. Chen, "A bibliometric analysis of research on carbon tax from 1989 to 2014," *Renewable & Sustainable Energy Reviews*, vol. 58, pp. 297–310, 2016.

[20] K. Randhawa, R. Wilden, and J. Hohberger, "A Bibliometric Review of Open Innovation: Setting a Research Agenda," *Journal of Product Innovation Management*, vol. 33, no. 6, pp. 750–772, 2016.

[21] A. Yataganbaba and I. Kurtbaş, "A scientific approach with bibliometric analysis related to brick and tile drying: A review," *Renewable & Sustainable Energy Reviews*, vol. 59, pp. 206–224, 2016.

[22] X. Chen, B. Chen, C. Zhang, and T. Hao, "Discovering the Recent Research in Natural Language Processing Field Based on a Statistical Approach," in *Emerging Technologies for Education*, vol. 10676 of *Lecture Notes in Computer Science*, pp. 507–517, Springer International Publishing, Cham, 2017.

[23] H. J. Kim, D. Y. Yoon, E. S. Kim, K. Lee, J. S. Bae, and J.-H. Lee, "The 100 most-cited articles in neuroimaging: A bibliometric analysis," *Results in Physics*, vol. 139, pp. 149–156, 2016.

[24] X. Chen, H. Weng, and T. Hao, "A Data-Driven Approach for Discovering the Recent Research Status of Diabetes in China," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 10594, pp. 89–101, 2017.

[25] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, vol. 46, supplement 1, pp. 234–240, 1970.

[26] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, 2002.

[27] D. R. Radev, M. T. Joseph, B. Gibson, and P. Muthukrishnan, "A bibliometric and network analysis of the field of computational linguistics," *Journal of the Association for Information Science and Technology*, vol. 67, no. 3, pp. 683–706, 2016.

[28] M. Grandjean, "A social network analysis of Twitter: Mapping the digital humanities community," *Cogent Arts and Humanities*, vol. 3, no. 1, Article ID 1171458, 2016.

[29] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, "A high-resolution human contact network for infectious disease transmission," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 107, no. 51, pp. 22020–22025, 2010.

[30] J. Scott, "Social network analysis," *Sage*, 2017.

[31] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[32] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proceedings of the the 43rd Annual Meeting*, pp. 363–370, Ann Arbor, Michigan, June 2005.

[33] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *American Association for the Advancement of Science: Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[34] A. F. El-Samak and W. Ashour, "Optimization of Traveling Salesman Problem Using Affinity Propagation Clustering and Genetic Algorithm," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 5, no. 4, 2015.

[35] G. Miao, N. Himayat, and G. Y. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Transactions on Communications*, vol. 58, no. 2, pp. 545–554, 2010.

[36] I. S. MacKenzie and R. W. Soukoreff, "Text entry for mobile computing: Models and methods, theory and practice," *Human–Computer Interaction*, vol. 17, no. 2-3, pp. 147–198, 2002.

[37] D. L. Strayer and F. A. Drews, "Cell-phone-induced driver distraction," *Current Directions in Psychological Science*, vol. 16, no. 3, pp. 128–131, 2007.

[38] J. Cao, T. Chen, and J. Fan, "Landmark recognition with compact BoW histogram and ensemble ELM," *Multimedia Tools and Applications*, 2015.

[39] M. M. Mostafa, "More than words: social networks' text mining for consumer brand sentiments," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4241–4251, 2013.

[40] H. Jiang, M. Qiang, and P. Lin, "A topic modeling based bibliometric exploration of hydropower research," *Renewable & Sustainable Energy Reviews*, vol. 57, pp. 226–237, 2016.

[41] H. B. Mann, "Nonparametric tests against trend," *Econometrica*, vol. 13, pp. 245–259, 1945.

[42] D. B. Lange and M. Oshima, "Seven Good Reasons for Mobile Agents," *Communications of the ACM*, vol. 42, no. 3, pp. 88-89, 1999.

[43] M. B. Mollah, M. A. K. Azad, and A. Vasilakos, "Security and privacy challenges in mobile cloud computing: Survey and way ahead," *Journal of Network and Computer Applications*, vol. 84, pp. 38–54, 2017.

[44] Y. Xi, L. Schwiebert, and W. Shi, "Privacy preserving shortest path routing with an application to navigation," *Pervasive and Mobile Computing*, vol. 13, pp. 142–149, 2014.

[45] T. Yan, V. Kumar, and D. Ganesan, "CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys '10)*, pp. 77–90, ACM, San Francisco, Calif, USA, June 2010.

[46] Y. Cai, S. Li, Y. Cheng, and R. Ji, "Local consistent hierarchical Hough Match for image re-ranking," *Journal of Visual Communication and Image Representation*, vol. 37, pp. 32–39, 2016.

[47] D. Eddington and D. Elzinga, "The phonetic context of american english flapping: Quantitative evidence," *Language and Speech*, vol. 51, no. 3, pp. 245–266, 2008.

[48] X. Wang, H. Zhang, S. Yuan, J. Wang, and Y. Zhou, "Sentiment processing of social media information from both wireless and wired network," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, article no. 164, 2016.

[49] Z. Wang, X. Cui, L. Gao, Q. Yin, L. Ke, and S. Zhang, "A hybrid model of sentimental entity recognition on mobile social media," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, article no. 253, 2016.

[50] K. E. C. Levy and M. Franklin, "Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry," *Social Science Computer Review*, vol. 32, no. 2, pp. 182–194, 2014.