Scientometrics
DOI 10.1007/s11192-015-1559-9

Mining research trends with anomaly detection models: the case of social computing research

Qing Cheng · Xin Lu · Zhong Liu · Jincai Huang

Received: 4 June 2014

Abstract We proposed in this study to use anomaly detection models to discover research trends. The application was illustrated by applying a rule-based anomaly detector (WSARE), which was typically used for biosurveillance purpose, in the research trend analysis in social computing research. Based on articles collected from SCI-EXPANDED and CPCI-S databases during 2000 to 2013, we found that the number of social computing studies went up significantly in the past decade, with computer science and engineering among the top important subjects. Followed by China, USA was the largest contributor for studies in this field. According to anomaly detected by the WSARE, social computing research gradually shifted from its traditional fields such as computer science and engineering, to the fields of medical and health, and communication, etc. There was an emerging of various new subjects in recent years, including sentimental analysis, crowdsourcing and e-health. We applied an interdisciplinary network evolution analysis to track changes in interdisciplinary collaboration, and found that most subject categories closely collaborate with subjects of computer science and engineering. Our study revealed that,

Q. Cheng (✉) · Z. Liu · J. Huang
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, People's Republic of China
e-mail: sgggps@163.com

X. Lu
College of Information System and Management, National University of Defense Technology, Changsha 410073, People's Republic of China

X. Lu
Flowminder Foundation, 17177 Stockholm, Sweden

X. Lu
Department of Public Health Sciences, Karolinska Institutet, 17177 Stockholm, Sweden

X. Lu
Division of Infectious Disease, Key Laboratory of Surveillance and Early-Warning on Infectious Disease, Chinese Centre for Disease Control and Prevention, Beijing 102206, People's Republic of China

Published online: 15 March 2015

anomaly detection models had high potentials in mining hidden research trends and may provided useful tools in the study of forecasting in other fields.

Keywords Research trend · Anomaly detection · WSARE · Social computing · Interdisciplinary networks

Introduction

In the era of information-explosion, the rapid development of science and technology requires researchers to be able to get immediate access to the latest research developments in related fields. To be updated with the latest research, researchers need to know what is changing in the research process and where the trend is going. Thereby, mining research trends from enormous publications has become increasingly important for both researchers and institutional policy makers. In the past years, mining research trends has attracted much attention from researchers across multiple domains and has been widely applied in various fields, such as computer science (Hoonlor et al. 2013), remote sensing (Zhuang et al. 2013), materials science (Kademani et al. 2013), electric vehicle (Hu et al. 2014) and engineering (Ucar et al. 2014). Moreover, companies and organizations such as Elsevier, Thomson Scientific and Google, are also making considerable effort in developing useful tools for analyzing research trends, such as Thomson Data Analyzer, Scopus, SciVal and Google trends. Previous approaches are mostly based on bibliometric methods and network analytics. By extracting "explicit" statistics and explaining relationships between studied variables, various visualization tools are then used for presenting analysis results intuitively. However, such approaches may fail to find "implicit" research trends, which means that they can rarely detect novel topics or events in a collection of literature and are difficult to answer what factors lead to changes in the specific field of study.

To overcome this limitation, we test a novel anomaly detection method to mine research trends by applying it to the field of social computing research. In doing so, we contribute to a better understanding of interdisciplinary collaboration in social computing research and find several novel topics which are emerging in recent year.

Related work

Approaches for research trends mining

Previous studies had applied many approaches to analyze research trends. The most typical, yet time consuming approach is to collect and review numerous literatures in the field, and summarize trends and directions for further research with qualitative studies (Wang et al. 2013). The second approach is called bibliometric method, which conduct statistical analysis of publication output, subject category, author, country and keywords, etc., such as keyword frequency and citation analysis (Wang et al. 2011; Sinha 2012; Fu et al. 2012; Liu et al. 2012). In recent years, network analysis was increasingly applied to mine hot topics and trace scientists research trends by analyzing relationships among keywords, country and research institute and author, including co-word analysis (Zhao and Zhang 2011), co-citation analysis (Lai and Wu 2005), co-authorship analysis (Glanzel

2000), etc. Nowadays, there have been some studies which integrated traditional bibliometric method and network analysis method to study research trends. For example, Hoonlor et al. (2013) analyzed the evolution of the computer science research landscape 1990–2010, Wang et al. (2014) characterized the key features and the evolution of social computing from two quantitative bibliometrical dimensions: statistics and topology. Additionally, various visualization technologies were proposed for presenting research trends, for example, CiteSpace and ArcGIS software had been used to demonstrate geographic distribution of authors or research institutes based on author address (Liu et al. 2011; Wang et al. 2012), ThemeRiver (Havre et al. 2002) and historiographic mapping (Garfield 2004) were used to show dynamic characteristics of studies in a field.

However, it is defective to evaluate research trends just by using bibliometric methods or network analysis methods. The bibliometric method focuses on simple statistics dimensions from a quantitative perspective, and can only get "explicit" statistics; The network analysis approach, on the other hand, focuses only on relationships between studied variables and thus fails to consider the specific content. As Shuai et al. (2012) suggested, it is not always true that citation data represent an explicit, objective expression of impact by scientists. And visualization technology only provides a better way to effectively convey information and to make the result to be more easily understandable. Thus, these methods fail to find "implicit" research trends. For example, they are difficult to answer what factors lead to changes in the specific field of study, and they rarely analyzed subjects with small number of existing studies, even the number of studies in these subjects may greatly increase in the future.

Anomaly detection methods

Anomaly detection methods were adopted to track research trends in order to overcome above limitations. The beginning of a new trend can be characterized as an innovative anomaly which is caused by certain changes in the research process (Basu and Meckesheimer 2007). For example, after a thorough investigation on the issue of anomalies in evaluative scientometrics, Glanzel (2013) and Glanzel and Moed (2013) point out that "While in many fields anomalies can simply be discarded as being exceptions, in bibliometrics the extreme values represent the high-end of research performance and therefore deserve special attention". A huge variety of anomaly detection methods have been developed by scholars from interdisciplinary communities, including physicists, computer scientists and mathematicians, etc. (Niu et al. 2011; Chandola et al. 2009), the latest representative research from aspects of statistical method, neural networks, clustering methods and other machine learning techniques.

Statistical methods fit a statistical model (usually for normal behavior) to the given data and then apply a statistical inference test to determine if an new instance belongs to this model or not. Instances that have a low probability to be generated from the learnt model, are declared as anomalies (Chandola et al. 2009). Earlier, *z*-index (Grubbs 1969) and box plot (Laurikkala et al. 2000) were used to determine which points are anomalies. More recently, Prathap (2014) showed that *z*-index is able to account for the high-end of research performance which is found as significant outliers in the tail of a citation distribution. Neural networks are typically used by training on normal data, then marking new instances as anomalies if they are rejected by the neural network or if they score high enough on some error measure (Chandola et al. 2009). Many variants of the basic neural network technique have been proposed to detect anomalies, such as restricted Boltzmann machine (Fiore et al. 2013), multilayer feedforward neural network and radial basis function network (Abuadlla et al. 2014). Clustering analysis applies a known clustering based algorithm to the data set and declare any data instance that does not belong to any cluster as anomalous (Chandola et al. 2009). Several clustering algorithms were used to detect various anomalies. For example, Srivastava and Zane-Ulman (2005) studied the *K*-means clustering, spectral clustering and von Mises Fisher (vMF) clustering to discover recurring anomalies on thousands of free-text reports. Recently, Wang et al. (2014) provided a novel framework of anomaly intrusion detection based on the Affinity Propagation (AP) clustering algorithm. In addition, there are other machine learning methods which have been widely used in abnormal findings, for example, nonlinear analysis (Palmieri and Fiore 2010), ensemble learning (Shoemaker and Hall 2011), multilevel immune learning (Dasgupta et al. 2005), two-stage machine learning (Palmieri et al. 2014), etc.

However, statistical methods are generally suited to quantitative data sets and are preferred for point anomalies detection; neural networks, on the other hand, requires labeled training data, which is typically difficult to collect if anomalies are rare; clustering analysis may perform badly if the data is not cluster-like or if the anomalous instances are frequent enough to form their own clusters. Furthermore, these methods are designed for discovering the anomalies for a single attribute. However, while isolated anomalies in attribute space were not always indicators of “novelty” research areas, the combination of features might indicate anomalies of particular research directions. For example, if studies with the keyword “smart cities” carried out in “Sweden” in the field of “computer science” (a combination of keywords, country and research area) increases unusually in the previous years, it might indicate a new trend.

To be able to detect anomalies formed by combination of arbitrary data attributes, we borrow theories from the field of biosurveillance and attempt to build a model which is capable of detecting anomalies not only on a single attribute but also on attribute combinations. The model, WASRE (What’s Strange About Recent Events), is the state-of-the-art model in biosurveillance for early disease outbreak detection. This model is not only able to find anomalies, but also capable of identifying identify anomalies that occur for groups with combinations of features (Wong et al. 2002, 2005). In this study, we showed how to apply WSARE in research trend analysis, and illustrated the application by studying social computing research field.

Social computing

Social computing represents a new computing paradigm and an interdisciplinary research and application field (Wang et al. 2007). It has become a hot topic attracting broad interest from not only researchers but also technologists, software and online game vendors, Web entrepreneurs, business strategists, political analysts, and digital government practitioners. Parameswaran and Whinston (2007) suggested that social computing should be a priority for researchers and business leaders and illustrated the fundamental shifts in communication, computing, collaboration, and commerce. Tracing research trends of social computing can help us to grasp the realtime development and future direction of science and technology.

The number of research papers published in social computing conferences and journals has increased rapidly in the past decade, but there are rare studies on patterns and trends for the field, and most studies have focused on challenges, directions, and landscapes in specific social computing fields and on specific social computing topics (Parameswaran and Whinston 2007; Pascu 2008; King et al. 2009; Xu et al. 2010; Chen et al. 2011), a comprehensive comment of social computing research has never been applied. In this study, we aim to fill in this gap in knowledge by combining anomaly detection model (WSARE) and the traditional bibliometric methods to investigate trends of social computing research in recent years.

Data sources and methodology

Data collection

We identified a set of keywords about social computing from two sources: the topics of "IEEE International Conference on Social Computing" as of 2011 (SocialCom 2011), which provided 16 topics in 3 years, and the description of social computing in Wikipedia (Wiki 2014), from which keywords were extracted from both Wikipedia and the most cited papers with "social computing" contained in the title for each topic. After excluding common terms such as "data mining", "web 2.0", etc., 17 keywords (including social intelligence, social simulation, social engineering, social informatics, social computing, social software, computational social science, mobile social, social media, human computation, social bookmark, folksonomy, wisdom of crowds, social tag, reality mining, user generated and crowdsourcing) were eventually identified to depict social computing research. We then retrieved articles using these keywords as search terms in two citation databases (Science Citation Index Expanded (SCI-EXPANDED) and Conference Proceedings Citation Index-Science (CPCI-S)) as of 2013.

After excluding duplicates, 6009 articles were identified as being in the social computing field (Fig. 1). Each article includes authors, title journal, language, document type keywords, abstract, author address reprint address, cited references and subject category etc. For simplicity, the address of each article was determined by the first address of the corresponding author. In this study, we analyze research trends of social computing field mainly from subject category, author address (country or state) and keywords.

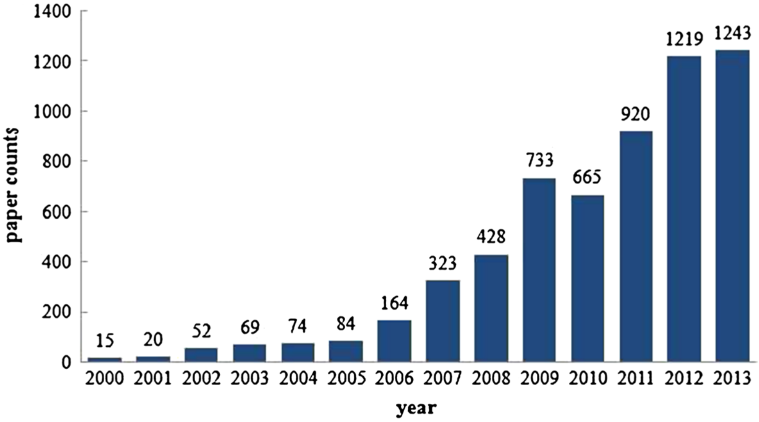


Fig. 1 Number of articles published in social computing-related field

Table 1 A sample two-by-two contingency table

	C_{recent}	C_{baseline}
Records match rule r	48	45
Records dont match rule r	86	220

Method

In this paper, we use a rule-based anomaly detector WSARE to detect anomalies, this method compares recent data against a baseline distribution with the aim of finding rules that summarize significant patterns of anomalies. Each rule is made up of components of the form $X_i = V_i^j$, where X_i is the ith attribute and V_i^j is the jth value of the attribute, Multiple components are joined together by a logical AND. for example, a two-component rule would be country=USA AND subject category = computer science (i.e. X_i = country, V_i^j = USA, and X_i = subject category, V_i^j = computer science). The description of WSARE is as follows

1. Create baseline data from a specified baseline period
   Suppose we treat all records within the current year as "recent" events and we would like to capture any current trends in the data. One solution would be to use only the most recent data, such as data from the previous year, thus, data from the last 5 years is used for the baseline period in this paper.
2. Score one-component rule and two-component rule
   For any one-component rule r, count records that match r from the dataset for recent and baseline period, C_{recent} and C_{baseline} respectively. A two-by-two contingency table if formed, as Table 1.
   Then, we use the Chi Squared test for independence of variables and get a p-value, which is referred to as the "score". However, since we are searching for anomalies, the counts in the contingency table frequently involve small numbers which violates the validity of Chi Squared test, in this case, we use Fisher's Exact Test to find the "score" for each rule. After testing all rules, the best one-component rule (BR^1) can be found, and then the "score" from the test is denoted as $Score(\mathrm{BR}^1)$.
   For two-component rule, the algorithm then attempts to find the best two-component rule for the year by adding on one extra component to BR^1 by a greedy search. This extra component is determined by supplementing BR^1 with all possible attribute-value pairs except for the one that is already in BR^1, and selecting the resulting two-component rule with the best score. Denote the best scoring rule found above as BR, Further details on the creation of n-component rules are introduced in (Wong et al. 2005).
3. Calculate P value by randomization test
   Denote the best scoring rule found in (2) or (3) as BR, for each iteration j (usually 1000), shuffle the dates between records in the recent and the baseline datasets to produce a randomized dataset called $\mathrm{DB}^j_{\text{rand}}$, then find the best scoring rule BR^j. The compensated P value CPV is calculated as

$$\mathrm{CPV} = \frac{\text{No. of times Score}(\mathrm{BR}^j) < \text{Score}(\mathrm{BR})}{\text{No. of randomaization test iterations}} \tag{1}$$

Finally, an anomaly is found for BR if CPV < 0.05.

Results and discussion

Emerging research trends analysis in author country and subject category

A total of 6009 articles were retrieved during 2000–2013, among these articles, 376 do not contain country information, 42 do not have a subject category. Based on the classification of subject categories in the 2010 Journal Citation Reports, social computing covered 127 subject categories. The top 10 subject categories were computer science (3796), engineering (1200), telecommunications (399), information science and library science (318), education and educational research (277), business and economics (219), operations research and management Science (202), health care sciences and services (183), psychology (154) and neurosciences and neurology (136) (Fig. 2). The number of articles in computer science and engineering has remained in the top two from 2000 to 2013. This showed the emphasis on social computing research in computer science and engineering.

Most social computing studies were from a few countries: the productivity ranking was headed by the USA (1728); and China (521) published the second highest number of articles; followed by England (321), Germany (311), Spain (234), South Korea (213), Canada (203), Australia (191), Japan (170) and Italy (166) (Fig. 3). 67.53 % of the studies were from the top 10 countries, where 28.76 and 8.67 % of the studies were form the USA and China respectively. Moreover, the number of articles in USA rose sharply from less than 50 in 2000 to about 450 in 2013 and the increase was more dramatic after 2011. Overall, the number of articles in China grew steadily and reached a peak at 144 in 2012, while in 2013 was there an obvious decline.

Taking both the author country and subject category of the articles into consideration, we use the WSARE (two-component rules: author country and subject category) to detect emerging research trends, a total of 31 anomalies were found during 2011–2013, with 19 anomaly increases and 12 anomaly decreases (Fig. 4).

During 2011–2012, we can see that there is a significant increase of social computing-related studies in China, in contrast to an anomaly decrease in England; Social computing-related studies in physics and operations research and management science appeared anomaly increase, while education and educational research and neurosciences and

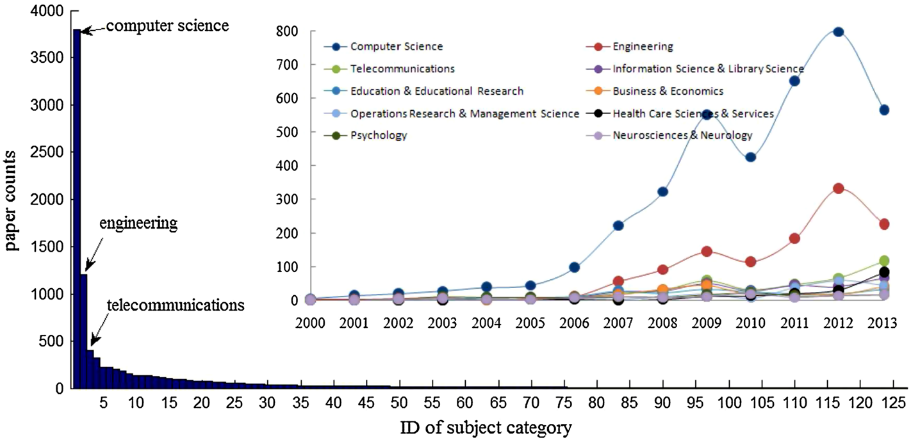


Fig. 2 Subject category distribution of social computing research subject categories

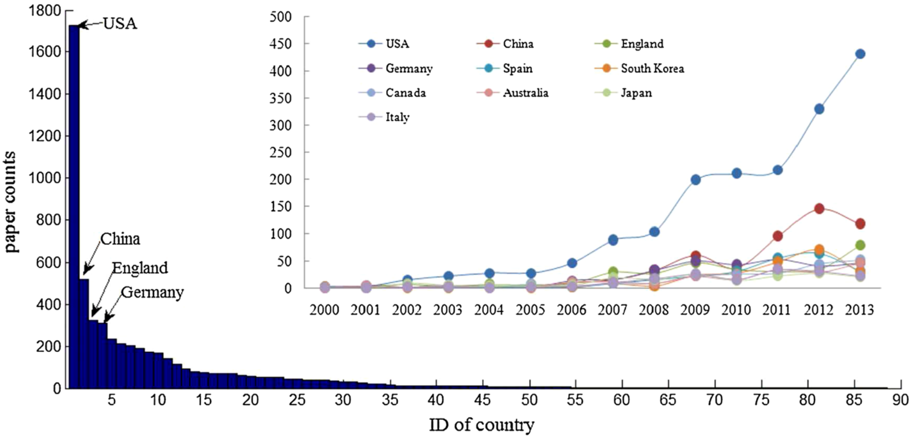


Fig. 3 Country distribution of social computing research

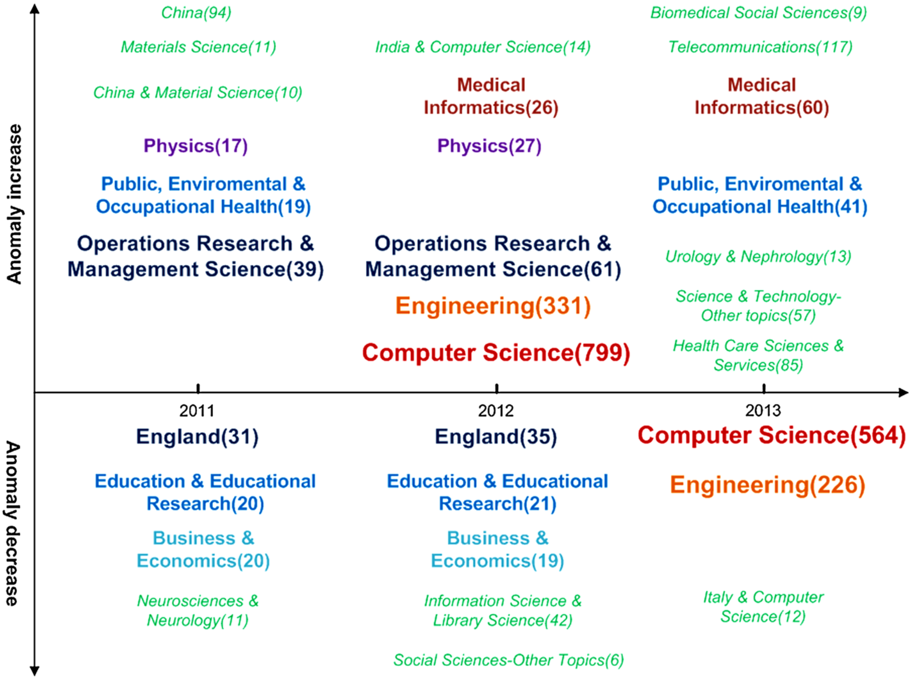

Anomaly increase
Anomaly decrease
China(94)
Materials Science(11)
China & Material Science(10)
Physics(17)
Public, Enviromental & Occupational Health(19)
Operations Research & Management Science(39)
India & Computer Science(14)
Medical Informatics(26)
Physics(27)
Operations Research & Management Science(61)
Engineering(331)
Computer Science(799)
Biomedical Social Sciences(9)
Telecommunications(117)
Medical Informatics(60)
Public, Enviromental & Occupational Health(41)
Urology & Nephrology(13)
Science & Technology-Other topics(57)
Health Care Sciences & Services(85)
2011
2012
2013
England(31)
Education & Educational Research(20)
Business & Economics(20)
Neurosciences & Neurology(11)
England(35)
Education & Educational Research(21)
Business & Economics(19)
Information Science & Library Science(42)
Social Sciences-Other Topics(6)
Computer Science(564)
Engineering(226)
Italy & Computer Science(12)

Fig. 4 Anomalies found by using rule-based anomaly pattern detector. *Numbers* indicate articles published in this category or country

neurology showed abnormal increase. In 2012, the studies of social computing grew rapidly, as shown in Fig. 1. This grow may be attributed to the notable growth of social computing-related studies in top subject categories, such as computer science, engineering and operations research and management science, because the abnormal increases were detected for social computing-related studies in subject categories of computer science, engineering and operations research and management science (Fig. 4), and most social computing-related studies are concentrated in these subject categories (Fig. 2). Nevertheless, from Fig. 4, we observed the anomaly decrease of information science and library science, business and economics, education and educational as subject categories in social computing research.

Compared to 2012, the number of articles in 2013 continued growing, but the growth rate was significantly reduced (Fig. 1). It is noteworthy that traditional fields of social computing research, such as computer science and engineering decreased significantly (Fig. 2), and Italy's social computing research in computer science also showed an anomaly decrease. Actually, the amount of social media data dealing with medical and health issues increased significantly recently (Agarwal 2011). For example, a 2013 online survey found out that 59.9 % patients used twitter for increasing knowledge and exchanging advice and 52.3 % used facebook for social support and exchanging advice (Antheunis et al. 2013). Furthermore, the number of related articles on health service and telecommunications was found anomaly increase, mainly including medical informatics, biomedical social sciences, healthcare sciences and services and urology and nephrology and telecommunications (Fig. 4). This showed that social computing research gradually began to shift from its traditional research field such as computer science and engineering, to other fields, such as health service and communications.

Keywords analysis

Keywords provide more detailed information on research trends (Zhang et al. 2010). In order to trace dynamic changes in the social computing field, we exact keywords from articles with valid keywords information (4116 out of the 6009 articles). Besides, we used xsimilarity package (xsimilarity 2014) (an approach of similar sentence retrieval based on edit distance) to preprocess keywords and merge keywords with the same meaning, for example, "recommender system" and "recommend systems" should be considered as identical keywords. The top 20 high-frequency keywords during 2000–2013 were visualized in Fig. 5, revealing the diversity and the richness of recent social computing topics.

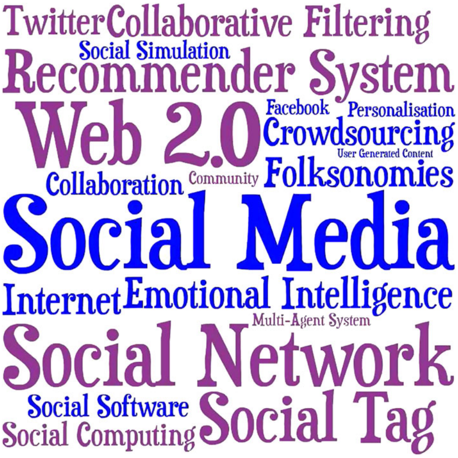


Fig. 5 Word cloud of keywords in social computing research. Font size is proportional to the frequency of words

Apparently, social media, social network, recommender system, web 2.0 and social tag, etc. are among the popular topics in social computing research.

By using WSARE (three-component rules: country, subject category and keywords), we did not found anomalies with two or three-component rules with the addition of keywords attribute. There are 32 one-component rule anomalies with keywords being the attribute (23 anomaly increases and 9 anomaly decreases), see Fig. 6.

According to anomaly detection analysis, "social media" was found anomaly increase during 2011–2013, while "social software", "web 2.0" appeared anomaly decrease. Since social computing went through three stages of groupware, social software and social media (Lugano 2012), it's no surprise that social computing studies focused primarily on social software a few years back, but social media analytics and related applications have a heavy presence in the social computing landscape recently, and researchers do tend to use "social media" more than "social software". In the past three years, the social computing research on social media platforms focused on "facebook", "twitter", because "twitter" showed a significant growth in 2011-2013 and "facebook" showed a significant growth in 2011 and 2013, as consistent with (Agarwal 2011), which also stated: "social tools such as facebook and twitter have become dominant drivers of future change in information and network technology along with the very functionality of modern society" (today, there are around one billion users on "facebook" and around 500 million users on "twitter"). Additionally, the keyword "microblog", which is also a social media platform, received a lot of attention in 2012, which is related to the growth trend of social computing research in China by considering the fact that the researchers in China tend to use "microblog", while in USA tend to use "twitter".

Furthermore, "collaborative filtering" is a technique used by some "Recommender systems" (Ricci et al. 2011) and they also increased significantly in 2011 and 2012, which

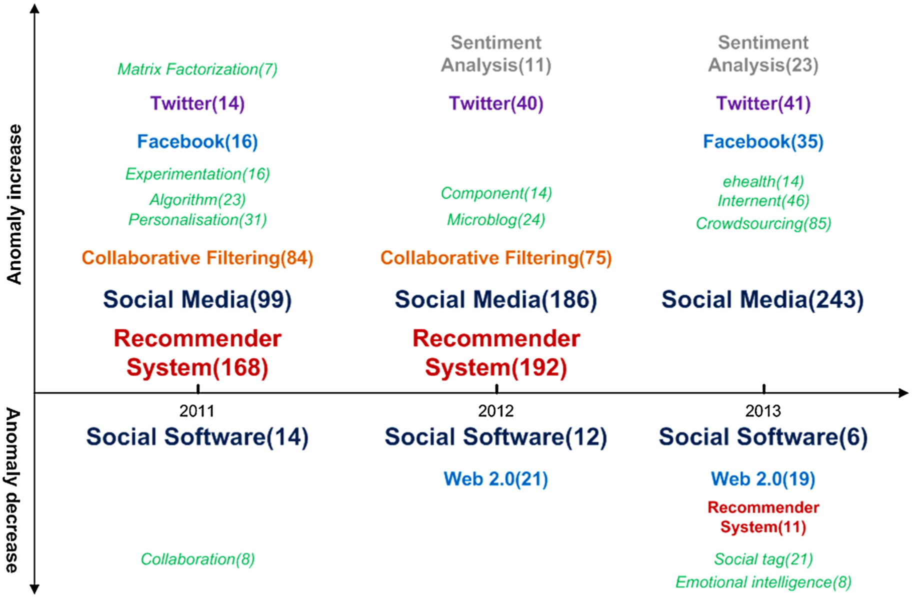


Fig. 6 Anomalies found by using rule-based anomaly pattern detector. *Numbers* indicate articles published in this category or country

suggested that the researches on "recommender system" has got more and more attention during 2011-2012. Recommender systems, as a kind of social computing service, have become extremely popular in recent years, and are applied in a variety of applications. It becomes a session topic for some well-known conferences such as RecSys, SIGIR, and KDD. However, the number of recommender systems studies decreased dramatically in 2013 (Fig. 6). Additionally, social computing has become more widely known because of its relationship to a number of recent trends. These include the growing popularity of "sentiment analysis", "crowdsourcing" and "e-health". "Sentiment analysis" was found an abnormal growth in 2013, this change is reflected in the fact that over 7000 articles have been written on the topic and Feldman (2013) also stated "there is a huge explosion today of 'sentiments' ". The notion of "crowdsourcing" was coined by Howe (2006) and the number of published crowdsourcing-related studies increased abnormally in 2013. This observation revealed its successful application in various fields and increasing attention for both academia and industry (Brabham et al. 2014; Carter et al. 2014; Faggiani et al. 2014). As Epirot Ludvik NeKaj, CEO and cofounder of Crowdsourcing Week, said: " Crowdsourcing is rapidly expanding as a disruptive practice in the working world and we see more and more companies trying to utilize the huge potential". According to Fig. 6, we can find that "e-health" is an anomaly in 2013, implying a potential novel research direction. Many successful applications are now witnessing a growing number of successful e-health developments in a wide range of domains, including health information networks, electronic health records, telemedicine services, wearable and portable monitoring systems, and health portals (Cruz-Cunla et al. 2010; Kierkegaard 2013). Besides, "e-health" was also listed as one of the UK's six vital new technology for future economic development by the UK Committee on Science and Technology (CST) in 2013.

Interdisciplinary network evolution analysis

Interdisciplinary collaboration is conceived as the main facilitator and driving force of the development of social computing, We focus on the changes of interdisciplinary collaboration on social computing research based on subject categories. This may either contribute to better understanding of the dynamics of disciplinary change, or may lead to disciplinary reconfiguration for facilitating development of social computing. The dynamic of interdisciplinary collaboration is characterized by a temporal network (called interdisciplinary network), with nodes being subject categories and edges being the co-occurrence of categories in the same article. Each edge is then weighted by the total number of co-occurrence between the connected categories. An overview network is visualized in Fig. 8, with thickness of line representing edges weight and node size representing nodes degree. Apparently, categories of computer science, engineering, environmental sciences and ecology and education and educational research have relatively large number of connections with others and are hubs in the network, indicating that collaborations in social computing research occurs often through these subject categories. Moreover, if we build interdisciplinary networks by year, the number of nodes, edges and the average degree, all sustained an upward trend (Fig. 7), which indicated the application of social computing has been increasingly wide and the interdisciplinary collaboration in social computing increased quickly as time went on.

Based on the interdisciplinary network by year on social computing, we find anomalies of interdisciplinary collaboration when interdisciplinary network changes with the use of WSARE. Firstly, we encode the interdisciplinary network into data items. Considering the fact that all networks evolve over time by addition and deletion of nodes and edges, which were reflected by modification of nodes degree and edges weight (for example, if a node is deleted, the nodes degree becomes 0, and, if an edge is deleted, the edges weight is 0). Suppose an edge connects a node v_1 and other node v_2 (the *id*s of v_1 and v_2 were n_1 and n_2 respectively), the edge was encoded as $(n_1 - 1) \times m + n_2 - 1$ called *edge-encoding number* (i.e. edge's *id*), which guarantees that each edge's edge-encoding number is unique, where m=127 since social computing covered 127 subject categories. Thus, the network is encoded into the data items, stored in the table *net2data*: $<time, n_1, n_2$, *edge-encoding number* $>$. For example, in 2013, a sub-network is a group of three nodes linked together, the nodes' *id* are 18, 63 and 64 respectively, and the edges' weight are 1, 1 and 5 respectively, as Fig. 9a, then, the sub-network can be encoded into data items as Fig. 9b. Therefore, a node degree is the number of occurrences of the node's *id*, and an edge weight is the number of occurrences of the *edge-encoding number*.

Secondly, by applying WSARE on data items (two-component rules: node's degree and edge's weight), we found 17 anomalies, all belong to edge anomaly (Fig. 10), among which 9 are anomaly increases (represented by red lines) and 8 are anomaly decreases (represented by green lines). The detected anomaly edges all belong to the computer

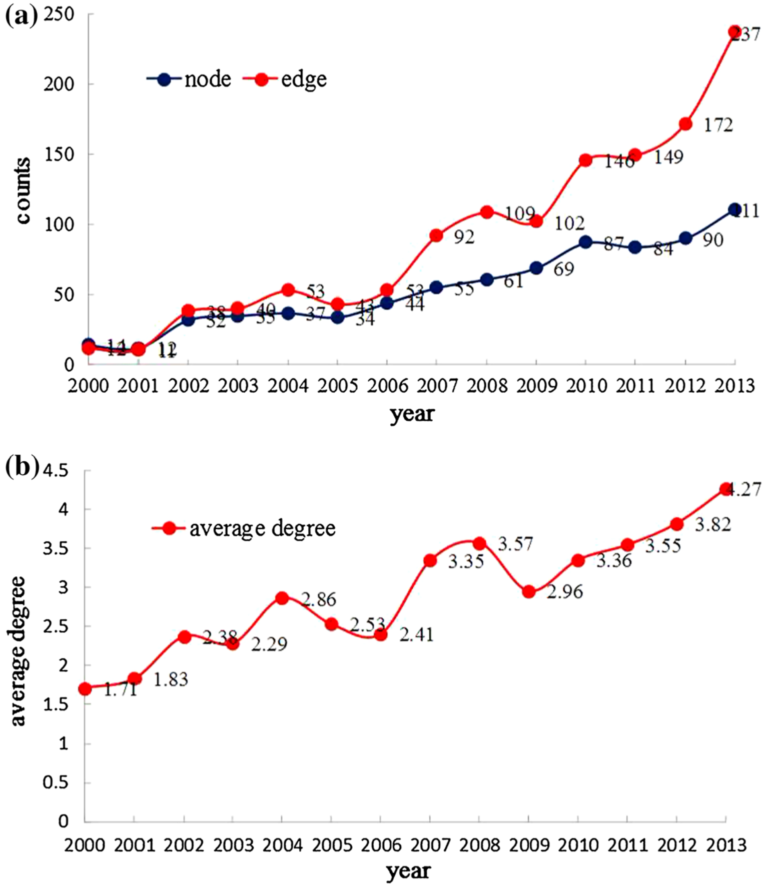


Fig. 7 The nodes and edges of the interdisciplinary networks

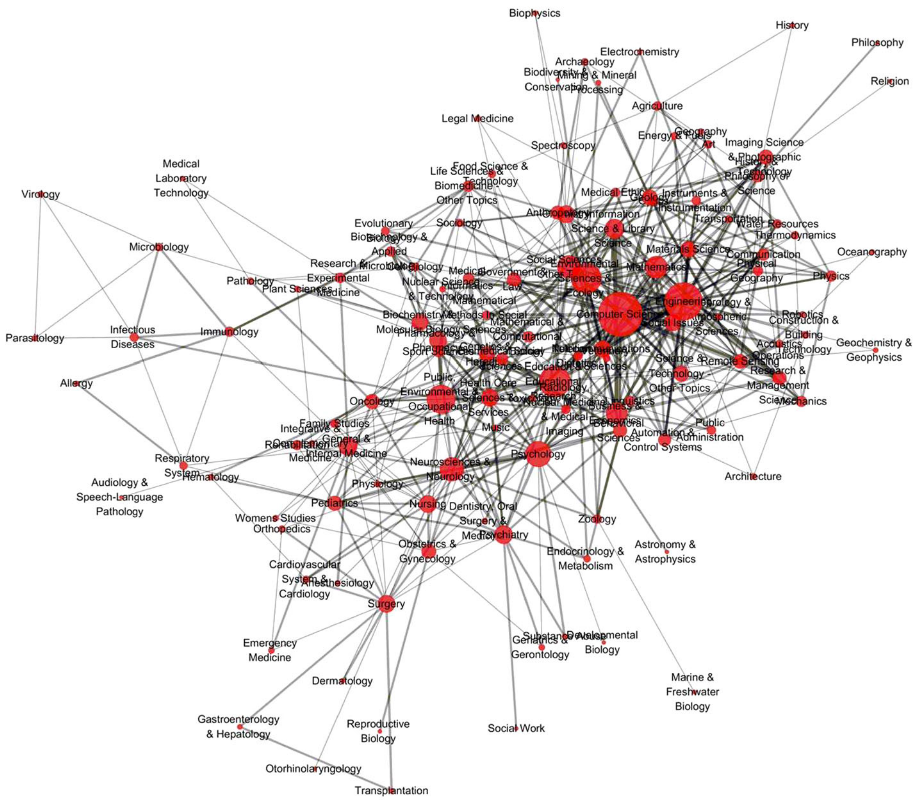
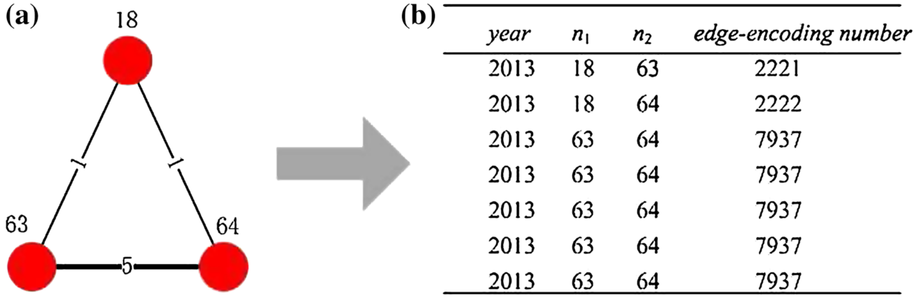

Biophysics
History
Philosophy
Religion
Electrochemistry
Archaeology
Biodiversity & Conservation
Mining & Mineral Processing
Agriculture
Legal Medicine
Spectroscopy
Energy & Fuels
Geography
Art
Imaging Science & Photographic Technology
History & Philosophy of Science
Medical Laboratory Technology
Virology
Food Science & Technology
Life Sciences & Biomedicine - Other Topics
Medical Ethics
Geology
Instruments & Instrumentation
Transportation
Water Resources
Thermodynamics
Anthropology
Information Science & Library Science
Sociology
Evolutionary Biology
Biotechnology & Applied Microbiology
Materials Science
Mathematics
Communication
Oceanography
Microbiology
Research & Experimental Medicine
Cell Biology
Social Sciences - Other Topics
Environmental Sciences & Ecology
Physical Geography
Physics
Pathology
Plant Sciences
Nuclear Science & Technology
Government & Law
Computer Science
Engineering
Meteorology & Atmospheric Sciences
Robotics
Construction & Building Technology
Parasitology
Infectious Diseases
Immunology
Biochemistry & Molecular Biology
Mathematical Methods In Social Sciences
Mathematical & Computational Biology
Social Issues
Acoustics
Operations Research & Management Science
Geochemistry & Geophysics
Pharmacology & Pharmacy
Genetics & Heredity
Sport Sciences
Telecommunications
Remote Sensing
Science & Technology - Other Topics
Allergy
Public, Environmental & Occupational Health
Health Care Sciences & Services
Education & Educational Research
Radiology, Nuclear Medicine & Medical Imaging
Linguistics
Business & Economics
Mechanics
Oncology
Toxicology
Behavioral Sciences
Automation & Control Systems
Public Administration
Integrative & Complementary Medicine
Family Studies
Rehabilitation
General & Internal Medicine
Music
Respiratory System
Hematology
Psychology
Audiology & Speech-Language Pathology
Neurosciences & Neurology
Architecture
Physiology
Pediatrics
Nursing
Dentistry, Oral Surgery & Medicine
Womens Studies
Orthopedics
Zoology
Psychiatry
Obstetrics & Gynecology
Astronomy & Astrophysics
Endocrinology & Metabolism
Cardiovascular System & Cardiology
Anesthesiology
Surgery
Emergency Medicine
Substance Abuse
Geriatrics & Gerontology
Developmental Biology
Marine & Freshwater Biology
Dermatology
Gastroenterology & Hepatology
Reproductive Biology
Social Work
Otorhinolaryngology
Transplantation

Fig. 8 Interdisciplinary network in 2000–2013, each *node* represents a subject category and article-sharing categories are connected through the edges of the graph, the *thickness* of which indicates the strength of connection



year	n_1	n_2	*edge-encoding number*
2013	18	63	2221
2013	18	64	2222
2013	63	64	7937
2013	63	64	7937
2013	63	64	7937
2013	63	64	7937
2013	63	64	7937

Fig. 9 A example to illustrate *net2data*

science centric-ego ne twork. Obviously, most anomaly edges are connected with the node of computer science or engineering. Specially, interdisciplinary collaboration in computer science and operations research and management science, engineering and operations

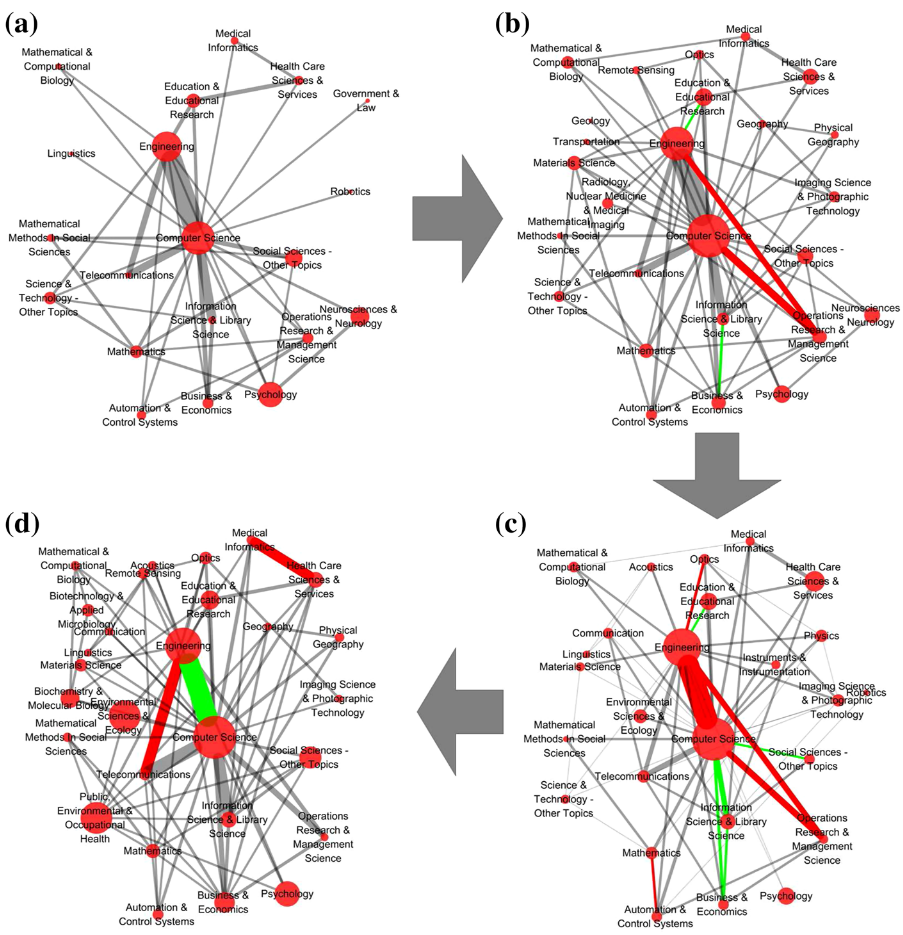

(a)
Mathematical & Computational Biology
Medical Informatics
Health Care Sciences & Services
Education & Educational Research
Government & Law
Linguistics
Engineering
Robotics
Mathematical Methods In Social Sciences
Computer Science
Social Sciences - Other Topics
Telecommunications
Science & Technology - Other Topics
Information Science & Library Science
Operations Research & Management Science
Neurosciences & Neurology
Mathematics
Business & Economics
Psychology
Automation & Control Systems
(b)
Medical Informatics
Mathematical & Computational Biology
Optics
Remote Sensing
Health Care Sciences & Services
Education & Educational Research
Geology
Geography
Physical Geography
Transportation
Engineering
Materials Science
Radiology, Nuclear Medicine & Medical Imaging
Imaging Science & Photographic Technology
Mathematical Methods In Social Sciences
Computer Science
Social Sciences - Other Topics
Telecommunications
Science & Technology - Other Topics
Information Science & Library Science
Neurosciences & Neurology
Operations Research & Management Science
Mathematics
Business & Economics
Psychology
Automation & Control Systems
(c)
Medical Informatics
Mathematical & Computational Biology
Acoustics
Optics
Health Care Sciences & Services
Education & Educational Research
Communication
Physics
Linguistics
Materials Science
Engineering
Instruments & Instrumentation
Imaging Science & Photographic Technology
Robotics
Environmental Sciences & Ecology
Mathematical Methods In Social Sciences
Computer Science
Social Sciences - Other Topics
Telecommunications
Science & Technology - Other Topics
Information Science & Library Science
Operations Research & Management Science
Mathematics
Business & Economics
Psychology
Automation & Control Systems
(d)
Medical Informatics
Mathematical & Computational Biology
Acoustics
Optics
Remote Sensing
Health Care Sciences & Services
Biotechnology & Applied Microbiology
Education & Educational Research
Communication
Geography
Physical Geography
Linguistics
Engineering
Materials Science
Imaging Science & Photographic Technology
Biochemistry & Molecular Biology
Environmental Sciences & Ecology
Mathematical Methods In Social Sciences
Computer Science
Social Sciences - Other Topics
Telecommunications
Public, Environmental & Occupational Health
Information Science & Library Science
Operations Research & Management Science
Mathematics
Business & Economics
Psychology
Automation & Control Systems

Fig. 10 Computer science-centric ego network evolution from 2011 to 2013, in which the *red line* represented abnormal increases, *green line* represented abnormal decreases. (Color figure online)

research and management science showed abnormal increase during 2011–2012. While the interdisciplinary collaboration in education and educational research and engineering, business and economics and information science and library science showed abnormal decline during 2011–2012. In addition, in 2012, although the social computing research in the field of computer science showed a significant growth, we found interdisciplinary cooperation in computer science and other subject categories abnormal decrease, such as information science and library science, social sciences-other topics, and business and economics. In 2013, we found interdisciplinary cooperation in engineering and telecommunications, medical informatics and health care science and services increased significantly (Fig. 10d), which suggested that the integration of such subject categories have been a main development tendency in social computing research.

Conclusion

By characterizing new research trends as innovative anomalies triggered by changes in research process, we have demonstrated the application of anomaly detection model in research trends analysis in this study. The application is illustrated by applying the biosurveillance model, WSARE, in the analysis of research trends for social computing during 2000 to 2013.

With a total of 6009 articles published in 127 subject categories, from a low of 15 in 2000 to a high of 1247 in 2013, the studies in social computing increased rapidly in the past 13 years. We find that computer science and engineering were the main subjects in social computing studies, and most of the publications were made by researchers from the USA and China. The model identifies a gradual shift of social computing research from its traditional subject categories, e.g., computer science and engineering, to subjects of health services and communication. Anomaly detection on keywords reveals that "facebook", "twitter" and "microblog" were the focus of social media platforms in social computing research. Even experienced a widely application during 2011–2012, the study of recommender system had decreased anomaly in the following year. Various technologies and applications were found increasing anomaly in 2013, including sentiment analysis, crowdsourcing and e-health, etc.

We further developed a framework for identifying anomalies in networks with subject categories being the nodes and co-appearance of subjects being the edges. The interdisciplinary network evolution analysis reveals that most subject categories closely collaborate with the traditional fields of social computing research, such as computer science, engineering, etc. The cooperation between engineering and telecommunications, medical informatics and health care science and services increased significantly in 2013.

Acknowledgments This research work was supported by the National Natural Science Foundation of China (No. 71301165, 91024030) and Hunan Provincial Innovation Foundation For Postgraduate under Grant No. CX2013B024. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Abuadlla, Y., Kvascev, G., Gajin, S., & Jovanovic, Z. (2014). Flow-based anomaly intrusion detection system using two neural network stages. *Computer Science and Information Systems*, *11*(2), 601–622.

Agarwal, N. (2011). Collective learning: An integrated use of social media in learning environment. In *Social media tools and platforms in learning environments* (pp. 37–51). Springer, Berlin.

Antheunis, M. L., Tates, K., & Nieboer, T. E. (2013). Patients' and health professionals' use of social media in health care: Motives, barriers and expectations. *Patient Education and Counseling*, *92*(3), 426–431.

Basu, S., & Meckesheimer, M. (2007). Automatic outlier detection for time series: An application to sensor data. *Knowledge and Information Systems*, *11*(2), 137–154.

Brabham, D. C., Ribisl, K. M., Kirchner, T. R., & Bernhardt, J. M. (2014). Crowdsourcing applications for public health. *American Journal of Preventive Medicine*, *46*(2), 179–187.

Carter, R. R., DiFeo, A., Bogie, K., Zhang, G.-Q., & Sun, J. (2014). Crowdsourcing awareness: Exploration of the ovarian cancer knowledge gap through Amazon Mechanical Turk. *PLOS ONE*, *9*(1). doi:10.1371/journal.pone.0085508

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–72.

Chen, Z., Yue, W., Shi, J., & Bu, X. (2011). A multi-agent based social computing collaboration selection approach in stable states. *Journal of Computational Information Systems*, *7*(16), 5785–5790.

Cruz-Cunla, M. M., Tavares, A. J., & Simoes, R. (2010). *Handbook of research on developments in e-health and telemedicine: Technological and social perspectives*. Hershey: Medical Information Science Reference.

Dasgupta, D., Yu, S., & Majumdar, N. (2005). MILA—Multilevel immune learning algorithm and its application to anomaly detection. *Soft Computing*, *9*(3), 172–184.

Faggiani, A., Gregori, E., Lenzini, L., Luoni, V., & Vecchio, A. (2014). Smartphone-based crowdsourcing for network monitoring: Opportunities, challenges, and a case study. *IEEE Communications Magazine*, *52*(1), 106–113.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82–89.

Fiore, U., Palmieri, F., Castiglione, A., & De Santis, A. (2013). Network anomaly detection with the restricted Boltzmann machine. *Neurocomputing*, *122*, 13–23.

Fu, J.-Y., Zhang, X., Zhao, Y.-H., Chen, D.-Z., & Huang, M.-H. (2012). Global performance of traditional Chinese medicine over three decades. *Scientometrics*, *90*(3), 945–958.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, *30*(2), 119–145.

Glanzel, W. (2000). Science in Scandinavia: A bibliometric approach. *Scientometrics*, *48*(2), 121–150.

Glanzel, W. (2013). High-end performance or outlier? Evaluating the tail of scientometric distributions. *Scientometrics*, *97*(1), 13–23.

Glanzel, W. & Moed, H. F. (2013). Opinion paper: Thoughts and facts on bibliometric indicators. *Scientometrics*, *96*(1), 381–394.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*, 1–21.

Havre, S., Hetzler, E., Whitney, P., & Nowell, L. (2002). ThemeRiver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 9–20.

Hoonlor, A., Szymanski, B. K., & Zaki, M. J. (2013). Trends in computer science research. *Communications of the ACM*, *56*(10), 74–83.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6), 1–4.

Hu, Y., Sun, J., Li, W., & Pan, Y. (2014). A scientometric study of global electric vehicle research. *Scientometrics*, *98*(2), 1269–1282.

Kademani, B. S., Sagar, A., Surwase, G., & Bhanumurthy, K. (2013). Publication trends in materials science: A global perspective. *Scientometrics*, *94*(3), 1275–1295.

Kierkegaard, P. (2013). eHealth in Denmark: A case study. *Journal of Medical Systems*, *37*(6). doi:10.1007/s10916-013-9991-y

King, I., Li, J., & Chan, K. T. (2009). A brief survey of computational approaches in social computing. In *IEEE international joint conference on neural networks (IJCNN)* (pp. 2699–2706).

Lai, K., & Wu, S. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information Processing & Management*, *41*(2), 313–330.

Laurikkala, J., Juhola, M., & Kentala, E. (2000). *Informal identification of outliers in medical data*. Berlin.

Liu, X., Zhang, L., & Hong, S. (2011). Global biodiversity research during 1900–2009: A bibliometric analysis. *Biodiversity and Conservation*, *20*(4), 807–826.

Liu, X., Zhan, F. B., Hong, S., Niu, B., & Liu, Y. (2012). A bibliometric study of earthquake research: 1900–2010. *Scientometrics*, *92*(3), 747–765.

Lugano, G. (2012). Social computing: A classification of existing paradigms. In *Proceedings—2012 ASE/IEEE international conference on privacy, security, risk and trust and 2012 ASE/IEEE international conference on social computing, SocialCom/PASSAT 2012* (pp. 377–382). Amsterdam.

Niu, Z., Shi, S., Sun, J., & He, X. (2011). A Survey of outlier detection methodologies and their applications. In *Artificial intelligence and computational intelligence, volume 7002 of lecture notes in artificial intelligence* (pp. 380–387).

Palmieri, F., & Fiore, U. (2010). Network anomaly detection through nonlinear analysis. *Computers & Security*, *29*(7), 737–755.

Palmieri, F., Fiore, U., & Castiglione, A. (2014). A distributed approach to network anomaly detection based on independent component analysis. *Concurrency and COmputation—Practice & Experience*, *26*(5), 1113–1129. doi:10.1002/cpe.3061

Parameswaran, M., & Whinston, A. B. (2007). Research issues in social computing. *Journal of the Association for Information Systems*, *8*(6), 336–350.

Pascu, C. (2008). *An empirical analysis of the creation, use and adoption of social computing applications*. Technical report, Institute for Prospective Technological Studies.

Prathap, G. (2014). Single parameter indices and bibliometric outliers. *Scientometrics*, *101*(3), 1781–1787.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook*, (pp. 1–35). Springer, New York.

Shoemaker, L., & Hall, L. (2011). Anomaly detection using ensembles. *Multiple classifier systems*. *Volume 6713 of lecture notes in computer science* (pp. 6–15). Springer, Berlin.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PLoS ONE*, *7*(11), doi:10.1371/journal.pone.0047523

Sinha, B. (2012). Global biopesticide research trends: A bibliometric assessment. *Indian Journal of Agricultural Sciences*, *82*(2), 95–101.

SocialCom. (2011). http://www.asesite.org/conferences/socialcom/2011/. Accessed 2011.

Srivastava, A. N., & Zane-Ulman, B. (2005). Discovering recurring anomalies in text reports regarding complex space systems. In *2005 IEEE aerospace conference* (vols. 1–4, pp. 3853–3862).

Ucar, I., Lopez-Fernandino, F., Rodriguez-Ulibarri, P., Sesma-Sanchez, L., Urrea-Mico, V., & Sevilla, J. (2014). Growth in the number of references in engineering journal papers during the 1972–2013 period. *Scientometrics*, *98*(3), 1855–1864.

Wang, F.-Y., Zeng, D., Carley, K. M., & Mao, W. (2007). Social computing: From social informatics to social intelligence. *IEEE Intelligent Systems*, *22*(2), 79–83.

Wang, H., He, Q., Liu, X., Zhuang, Y., & Hong, S. (2012). Global urbanization research from 1991 to 2009: A systematic research review. *Landscape and Urban Planning*, *104*(3–4), 299–309.

Wang, M.-H., Li, J., & Ho, Y.-S. (2011). Research articles published in water resources journals: A bibliometric analysis. *Desalination and Water Treatment*, *28*(1–3), 353–365.

Wang, T., Liu, Z., Xiu, B., Mo, H., & Zhang, Q. (2014). Characterizing the evolution of social computing research. *IEEE Intelligent Systems*, *29*(5), 48–56.

Wang, W., Guyet, T., Quiniou, R., Cordier, M.-O., Masseglia, F., & Zhang, X. (2014). Autonomic intrusion detection: Adaptively detecting anomalies over unlabeled audit data streams in computer networksy. *Knowledge-based system*, *70*, 103–117.

Wang, X., Wang, Z., & Xu, S. (2013). Tracing scientist's research trends realtimely. *Scientometrics*, *95*(2), 717–729.

Wiki. (2014). http://en.wikipedia.org/wiki/social_computing. Accessed 2014.

Wong, W., Moore, A., Cooper, G., & Wagner, M. (2002) Rule-based anomaly pattern detection for detecting disease outbreaks. In: *Eighteenth national conference on artificial intelligence (AAAI-02)/fourteenth innovative applications of artificial intelligence conference (IAAI-02), proceedings* (pp. 217–223).

Wong, W., Moore, A., Cooper, G., & Wagner, M. (2005). What's strange about recent events (WSARE): An algorithm for the early detection of disease outbreaks. *Journal of Machine Learning Research*, *6*, 1961–1998.

xsimilarity. (2014). https://code.google.com/p/xsimilarity/w/list. Accessed 2014.

Xu, Y., Luo, T., & He, H. (2010). Social computing research map. In *Proceedings - 2010 IEEE 2nd symposium on web society, SWS 2010* (pp. 158–164). Beijing.

Zhang, L., Wang, M.-H., Hu, J., & Ho, Y.-S. (2010). A review of published wetland research, 1991–2008: Ecological engineering and ecosystem restoration. *Ecological Engineering*, *36*(8), 973–980.

Zhao, L., & Zhang, Q. (2011). Mapping knowledge domains of Chinese digital library research output, 1994–2010. *Scientometrics*, *89*(1), 51–87.

Zhuang, Y., Liu, X., Nguyen, T., He, Q., & Hong, S. (2013). Global remote sensing research trends during 1991–2010: A bibliometric analysis. *Scientometrics*, *96*(1), 203–219.