Earth Sci Inform (2014) 7:147–152
DOI 10.1007/s12145-013-0132-4

REVIEW ARTICLE

Applications and researches of geographic information system technologies in bibliometrics

Wang Xuemei · Ma Mingguo · Li Xin · Zhang Zhiqiang

Received: 27 May 2013 / Accepted: 3 September 2013 / Published online: 18 October 2013

Abstract Although the research and application of a geographic information system (GIS) in bibliometrics remains in its initial stage, several valuable attempts have been made in recent years. This paper provides our overview regarding this area. We first reviewed the spatial information mining derived from literature, including structured and unstructured data. The spatial display and the basic spatial operations for the geographic information derived from literature were then introduced, demonstrating that GIS can be directly used to construct digital libraries. Some literature database websites have begun to utilize WEBGIS to display the spatial distribution of an author's location. Additionally, the spatial distribution information can be displayed in various modes with other specialized tools. Potential spatial analyses in bibliometrics were then discussed, introducing geostatistical and buffer analyses as case studies. Finally, several bibliometric indicators attached with research units were investigated. When the quantitative research index units are linked with the research spatial position, they can be displayed, queried and retrieved spatially. Future work to advance the application and research of GIS in bibliometrics is still warranted.

Communicated by: H. A. Babaie

Foundation item Under the Special Project of Literature Ability of the Chinese Academy of Sciences and the National Natural Science Foundation of China (grant no.: 40701133).

W. Xuemei · M. Mingguo (✉) · L. Xin
Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou 730000, China
e-mail: mmg@lzb.ac.cn

W. Xuemei · Z. Zhiqiang
Lanzhou Branch of the National Science Library, Scientific Information Center for Resources and Environment, Chinese Academy of Sciences, Lanzhou 730000, China

Keywords Bibliometrics · Geographic information system (GIS) · Spatial visualization · Spatial analysis

Introduction

As a subdiscipline of information science, bibliometrics was first suggested by Alan Pritchard in 1969. Pritchard defined bibliometrics as "the application of mathematics and statistical methods to books and other media of communication" (Pritchard 1969). Bibliometrics has also been defined as a set of methods to quantitatively analyze scientific and technological literature (Bellis 2009). Therefore, bibliometrics is an interdisciplinary field that combines mathematics, statistics and information science to study literature mining. Following the tremendous advancements in computer and networking technology, bibliometrics has developed in terms of visualization, networking and quantified factors (Börner et al. 2003; Van and Waltman 2010; Zhao and Xu 2010).

The literature normally involves some spatial-related information. The basic information sources are the authors' institutes, cities and countries of articles, citations and cited articles. Toponyms are names used to refer to locations, which can be linked with the points or regions of actual geographic coordinates. A set of toponym information is involved in the articles, which can be extracted based on the toponym resolution (Allison and Christian 1994). Furthermore, the literature concerned with geosciences generally involves more spatial-related information, including the study areas, sampling points, observation points, sampling units and sampling strips.

A geographic information system (GIS) integrates the hardware, software and data for capturing, managing, analyzing and displaying all forms of geographically referenced information. GIS is widely used in many professions. Preliminary applications and initial research have developed in recent years

(Frenken et al. 2009). Most of these works concern the spatial display of the spatial distribution of research and citations.

The spatially linked information can be managed, displayed and analyze by using GIS technologies. The interdisciplinary between bibliometrics and GIS is a newly developed direction (Allen 2001). In this paper, the advancement of applications and research of GIS technologies in bibliometrics was reviewed. Then, potential future research was discussed.

Spatial information mining from the literature

For spatial analyses and displays, raw spatial information must be extracted from a large mass of literature. Some information with the regular structure can be automatically and rapidly extracted. For example, the affiliations, cities and countries of the authors and co-authors are generally included in the literature. Because the names of cities and countries have fixed formats, they can be linked with the vector data within the spatial distribution information (Wang and Ma 2009). The affiliations are normally required to use the fixed formats by the journals. There are small differences among these journals. The shorter form or subsidiaries are sometimes used, which can be normalized based on the unit dictionary.

Abundant toponym (place name) information exists in literature abstracts and texts. These toponym data can also be linked with the vector data. However, the toponym data are very complicated. Various place forms, such as towns, mountains, rivers, lakes and streets, exist. Place name recognition is a significant part of text mining (Leidner 2007). Woodruff and Paunt (1994) have developed an algorithm, GIPSY, to automatically extract geographic coordinate index terms from text. GIPSY can index and retrieve the georeferenced document based on the US Geological Survey's Geographic Names Information System (GNIS). Several approaches for enhancing automatic place name recognition, disambiguation and visualization have been developed in recent years (Pouliquen et al. 2006; Bensalem and Kholladi 2010). Generally, toponym recognition can obtain >80 % accuracy. However, the least difficult scenario is the linkage between the place names and map coordinates, referencing place points or regions. Therefore, much of the research need focus on the thematic place names.

The study areas, sampling trips and sampling points are significant pieces of spatial location information in the geosciences literature. These spatial locations are sometimes represented as special place names but are usually represented by the coordinate information with latitude and longitude positions. The coordinate information can be translated into the vector layers of points, lines or polygons. These data can then be used to analyze the spatial distribution characteristics and research trends. The geographic coordinate information has a finite number of formats, which can also be automatically extracted from the literature. However, because little research regarding geographic coordinate information recognition exists, it can serve as a place name recognition method.

Spatial display and basic spatial operations

Spatial display and spatial query are the basic operations of GIS tools. A GIS can directly be used to construct digital libraries and store and manage map information, manage the spatial distribution patterns of the library resources and construct digital library networks when combined with Internet technologies (Liu et al. 2008). The spatial distribution of library resources was visually displayed on digital maps, enabling readers a sense of the space states of literature. The digital maps also help readers determine the positions of library resources. Barbara (1998) recognized the potential of using GIS in the library. The ARL GIS Literacy Project was initiated in 1992 as a multiphased project in partnership with ESRI (the leading producer of GIS software) and other public and private partners. As a result of this project, 89 % of 121 project participants provide GIS services (French 1999). A review was summarized on the library service using GIS, which mainly introduced two ways: to analyze service area populations and to manage facilities (Bishop and Mandel 2010). There are very rich GIS softwares and databases nowadays. The above functions can be realized by most of them. The ESRI AcrGIS is the most commonly used tool in our studies.

With the development of the WEBGIS, increasing numbers of literature database websites have begun to use WEBGIS to display the spatial distribution and the relationships between authors, co-authors and citations. For example, the BioMedExperts use WEBGIS to display the spatial distribution and geographic relationships of biomedical authors with their co-authors (www.biomedexperts.com/). When a biomedical author is indexed, all of the co-authors' cities and publications are displayed on a world map. The AuthorMapper was developed by a Springer-based database using Google Earth. The spatial positions of the authors are linked with the Springer literature database (www.authormapper.com/). When articles are indexed by different query modes, all of the results will be displayed according to the city distribution on a Google Map. Using the above tools, the network of relationships among addresses in scientific publications can also be displayed on the geographic map (Chen et al. 2008; Wang and Ma 2009; Olechnicka and Ploszaj 2010; Leydesdorff and Persson 2010).

Other browsing tools have also been developed for the geographic digital library. For example, the GeoVIBE combines a geographical model and a vector space model to form document representations that are tightly linked (Cai 2002). The Litmap was created to enable humanities scholars to read literature spatially (Barbara 2010). The Science of Science (Sci2) Tool is a modular toolset specifically designed for the

study of science (http://sci2.cns.iu.edu). It supports the temporal, geospatial, topical, and network analysis and visualization of scholarly datasets at the micro (individual), meso (local), and macro (global) levels (Börner 2011). The project of Alexandria Digital Library (ADL) was carried out from 1994 to 1999 and the Alexandria Digital Earth Prototype (ADEPT) was carried out from 1999 to 2005. The ADL developed the distributed digital library with collections of georeferenced materials and services across the Internet (Smith T R et al. 1996). The ADEPT project addresses issues of access, browsing, delivery and understanding of georeferenced library items. The visualization approaches include interfaces between some existing digital earth systems and the digital library, and spatial but nongeoreferenced information spaces (Ancona et al. 2002).

Nowadays, more and more journals and libraries begin to use the above tools to display the spatial distribution information of the authors or other research units. For example, the city addresses appearing in the high-impact journals Science, Nature, and PANS were analyzed. The published numbers of the cities can be browsed and indexed on the Google Map (www.nature.com/news/specials/cities/best-cities.html). The other case, the WEBGIS is used to show the published numbers of each institution. The circle sizes and colors represent the quantity and quality of the articles (www.excellencemapping.net).

The literature citation of a research unit (e.g., researcher, laboratory, institute/university, city, and country) can also be displayed on a spatial map. The spatial distribution of the cited authors displays the scientific impact of the research unit. The author, co-author and citation information has a relatively regular structure and can be automatically and rapidly indexed. This information is dynamically updated with the literature database in the above mode. Therefore, the statistics results linked with the spatial positions are also dynamically displayed on the Internet.

The interannual changes can be shown using a histogram or a trend line for each spatial vector feature. When the spatial feature is queried, the linked database of multi-year data can be presented (Chen et al. 2008; Ma et al. 2011).

A patent is also a significant literature format. The USPTO's patent data are mapped and displayed using overlays with Google Maps. The overlays indicate both the quantity and quality of patents at the city level (Leydesdorff and Bornmann 2012).

In contrast, displaying spatial-linked information with an irregular structure on the Internet is difficult. The most important step is the information extraction of other items from the irregular structure, which is of future focus. After this step, the spatial information linked with the vector data were extracted from the literature and could be inputted into the GIS tools. The spatial distribution information can then be displayed in various modes. For publication numbers, both points with graduated sizes and a polygon filling with graduated colors can be applied to show the quantitative difference. For cooperation articles or citation articles, the line links two countries that have cooperative research. The line thickness shows the cooperation paper number (Wang et al. 2012). For the 3D mode, the Red Army Long March was illustrated using dynamic a 3D GIS-based model, which allows users to interact and explore the 3D scene and consequently obtain useful information (Ma and Meng 2009).

Spatial analysis

Spatial analysis is the process of conducting various spatial data processing methods and obtaining information, clues and knowledge from the results. Here, spatial analysis can advance the deep data mining of the spatial-linked information from the literature. GIS provides specialized platforms for computing spatial relationships among spatial units. The most commonly used spatial analysis methods include tracking, buffer, overlay, route, network and geostatistical analyses and spatial interpolation. The application of GIS spatial analysis to bibliometrics is still in its initial stages.

Based on Scopus data, highly cited, recently published papers were displayed on Google Maps. The field-specific excellence can be identified and agglomerated in regions and cities (Bornmann et al. 2011; Bornmann and Leydesdorff 2011, 2012). The spatial interpolation method can be used to locate the high-density regions based on the highly cited paper numbers and can also be used to detect the "hot regions" of the scientific activities (Bornmann and Waltman 2011).

Geostatistical analyses were used in the geostatistics literature indexed from Web of Science, Scopus and Google Scholar (Heng et al. 2009). The focus regarded the citation rate analyses. A published research article density map was generated using the isotropic Gaussian kernel with a standard deviation of 0.5 arc degrees and using the CRs for each article separately. A correlation analysis between the density map and socio-economic variables was performed at the spatial scale to explore the driving factors. The socio-economic variables included night light images, environmental conditions, economic growth and population density, which are represented in the form of spatial distribution data and derived from remote sensing observation or spatial interpolation of socio-economic statistics. The spatial analyses used not only the geographically referenced data derived from the literature but also the relevant spatial data derived from other resources.

The buffer regions were used to analyze the spatial distribution of the sampling points in the Qinghai-Tibet Plateau as the distance with the traffic lines increased, such as railway and highway (Wang et al. 2012). The results indicated that the traffic lines were the most significant factors for distributing the sampling and observation points. The results also indicated that the point numbers decrease quickly with increasing distance to the traffic lines.

The spatial analysis of the geographic position information involved in the literature has the scale dependence, which mainly embodied with the statistics and analysis units. Sometimes we do not have enough data to find the spatial distribution regulars at a certain scale. But they can be revealed at the other scale. The fine scale geography may lead to more sophisticated indicators of scientific output (Carvalho and Batty 2006).

GIS-linked bibliometrics indicators

Indicators are the essence of bibliometrics, which were developed to quantify literature characteristics and amounts (Vinkler 2010). When these quantitative indicators are summarized or transformed for the research units, they can be linked with the spatial positions of these research units. Then they can be displayed, queried and retrieved spatially. There are two typical bibliometrics indicators. Firstly, journal citation measures are designed to assess significance and performance of individual journals. The impact factor is perhaps the most popular bibliometric product used in bibliometrics itself (Glänzel 2002). It is a quantitative index that reflects the average number of recently cited articles published in a journal (Garfield 1998). The journal citation measures cannot be linked a spatial feature directly. But the total impact factors, average impact factors and maximum impact factor can be summarized for a research unit during a given period. Secondly, some indicators are designed for quantify the real citation of some published source. The H-index, which is widely utilized, was developed to quantify an individual's scientific research output (Hirsch 2005). Since then, some variants of the H-index have been discussed (Schreiber 2008; Bornmann et al. 2008). These kinds of indicators are designed for the research units directly, therefore they can be linked a spatial feature directly.

Even though impact factor is commonly-used indicator of journal till now, several weaknesses of impact factor were thought of as existing (Vanclay 2012). Some improvements were made to evaluate outputs from scientific research. For example, the San Francisco Declaration on Research Assessment (DORA), initiated by the American Society for Cell Biology (ASCB) together with a group of editors and publishers of scholarly journals, encourages individuals and organizations to develop new scientometric indicators. Furthermore, the outputs from scientific research are not limited to the articles of journals or proceedings. More knowledge formats should be considered, such as reports, data, reagents, and software. Another example is the SCImago Institutions Rankings (SIR), which is a Research Evaluation Platform and Ranking Generator to analyze research outputs of universities and research-focused institutions. The SIR tries to develop journal and country scientific indicators. These indicators can be used to assess and analyze scientific domains (Bornmann et al. 2012). The above newly-developed scientometric indicators are normally designed for the funding agencies, institutions, publishers, researchers, and so on. These statistical units can also be linked spatial features and then be displayed by GIS.

The citation indicators of the papers or patents are one of most important formats of the knowledge flow among the research units, which have the direction property. They could be expressed by GIS tools visually when the direction property was considered. However, the above indicators represent all of the quantified estimations on the single direction of knowledge outflow. In contrast, the International Scholarly Impact of Scientific Research (ISISR) is designed to measure how often authors of a country are cited by researchers from outside the country in a given subject area, which is a quantified estimation of the single direction of knowledge inflow (Hassan and Haddawy 2013). The distance factor is also used to measure the spatial distribution pattern of the bidirectional knowledge flow. The spatial distance is calculated among the citing or cited papers based on GoogleMaps API and Yahoo! PlaceFinder (Wu 2013; Ahlgren et al. 2013). If the indicators had both geographical position and flow direction characteristics, they would be more suitable for spatial representation and analyses using GIS technologies. However, the relative research is still preliminary.

Conclusions and discussion

Interdisciplinary research is a developing trend in science and technology, and the application of GIS technologies in bibliometrics is an example of interdisciplinary research. But it still at his initial stage according to the above review. We think that it would become increasingly popular if more efforts were made in the future. Similar to scientific literature retrieval, we primarily search and copy from the paper journals that date from the end of the last century. However, electronic versions of journals have become very convenient now, which promote the knowledge flow of the citation largely. Therefore, only novel technologies are used in bibliometrics, the literature databases can serve as more functional and visual tools for authors and users.

To promote the application and research of GIS in bibliometrics, the following future research is warranted:

(1) Application of regular information structures. Regular information structures can be queried and calculated instantly and automatically. If more geographically relative information with regular structure exists in articles, the literature database can easily display its resources using GIS platforms. International standards can be set up, especially for scientific publication. The authors and their affiliation information must be standardized with regards to format and word use. The spatial position

involved in main text can also be standardized with unified templates. However, the written requirements of journals differ significantly.

(2) Preparation of basic thematic maps. For the researchers engaged in bibliometrics, collecting and arranging basic thematic maps for linking geographically relative information remains challenging. Although WEBGIS resources (e.g., Google Map and Yahoo! PlaceFinder) can be openly used, certain special applications require more preparation work. For a given study area (e.g., Qinghai-Tibet Plateau and Amazon Basin), a number of hot spot position names for local mountains, lakes, rivers, villages and so on exist. These position names are generally not included in the general map platforms, which require more collection and arrangement.

(3) Development of information recognition tools. The extraction of geographically relative information from numerous scientific articles using GIS in bibliometrics is a key process. Special tools to easily realize information recognition are rare. These tools should be developed for thematic information recognition. For example, lakes, roads and mountains are thematic layers and have relatively regular naming rules. A tool to extract the coordinate information with latitude and longitude position needs to be developed and output to the standard geographic format.

(4) Development of Internet display platforms. Many electronic maps online are widely used and have been rapidly developed. We can use these platforms to expand the functions of the information display and querying of spatial literature information. The BioMedExperts and Springer are successful case studies. However, these applications are simple. More functions should be released, especially for the application of GIS in literature representation. If possible, a specifically designed Internet display platform should be developed for realizing this application requirement.

References

Ahlgren P, Persson O, Tijssen R (2013) Geographical distance in bibliometric relations within epistemic communities. Scientometrics 95:771–784

Allen RS (2001) Interdisciplinary research: a literature-based examination of disciplinary intersections using a common tool, geographic information system (GIS). Sci Technol Libr 21(3–4):191–209

Allison W, Christian P (1994) GIPSY: automated geographic indexing of text documents. J Am Soc Inf Sci 45:645–655

Ancona D, Freeston M, Smith T, Fabrikant S (2002) Visual explorations for the Alexandria digital earth prototype. Börner K and Chen C (Eds.): visual Interfaces to digital libraries, LNCS 2539, pp. 199–213

Barbara PB (1998) Looking forwards: geographic information services and library in the future. Cartogr Geogr Inf Syst 25:161–171

Barbara LH (2010) Global networks in W.G. Sebald's Rings of Saturn. In The Undiscover'd Country: W.G. Sebald and the Poetics of Travel (ed.) Markus Zisselsberger. New York

Bellis DN (2009) Bibliometrics and citation analysis: from the Science citation index to cybermetrics. Scarecrow Press, pp. 417

Bensalem I, Kholladi MK (2010) Toponym disambiguation by arborescent relationships. J Comput Sci 6:653–659

Bishop BW, Mandel LH (2010) Utilizing geographic information systems (GIS) in library research. Libr Hi Tech 28(4):536–547

Börner K (2011) Plug-and-play macroscopes. Commun ACM 54(3):60–69

Börner K, Chen C, Boyack KW (2003) Visualizing knowledge domains. Annu Rev Inf Sci Technol 37:179–255

Bornmann L, Leydesdorff L (2011) Which cities produce worldwide excellent papers more than expected? A new mapping approach—using Google Maps—based on statistical significance testing. J Am Soc Inf Sci Technol 62:1954–1962

Bornmann L, Leydesdorff L (2012) Which cities' paper output and citation impact are above expectation in information science? Some improvements of our previous mapping approaches. J Informetrics 6:336–345

Bornmann L, Waltman L (2011) The detection of "hot regions" in the geography of science—A visualization approach by using density maps. J Informetrics 5:547–553

Bornmann L, Mutz R, Daniel HD (2008) Are there better indices for evaluation purposes than the h-index A comparison of nine different variants of the h index using data from biomedicine. J Am Soc Inf Sci Technol 59:830–837

Bornmann L, Leydesdorff L, Walch-Solimena C, Ettl C (2011) Mapping excellence in the geography of the sciences: an approach made possible by using Scopus data. J Informetrics 5:537–546

Bornmann L, Moya-Anegón F, Leydesdorff L (2012) The new excellence indicator in the world report of the SCImago institutions rankings 2011. J Informetrics 6(2):333–335

Cai G (2002) GeoVIBE: a visual interface to geographic digital library. Proceedings of Visual Interfaces to Digital Libraries, pp, 171–187, Portland

Carvalho R, Batty M (2006) The geography of scientific productivity: scaling in US computer science. J Stat Mech: Theor Exp 2006(10):P10012

Chen C, Song IY, Yuan X, Zhang J (2008) The thematic and citation landscape of data and knowledge engineering (1985–2007). Data Knowl Eng 67:234–259

French M (1999) The ARL GIS literacy project: support for government data services in the digital library. From http://www.iassistdata.org/downloads/iqvol241french.pdf

Frenken K, Hardeman S, Hoekman J (2009) Spatial scientometrics: towards a cumulative research program. J Informetrics 3:222–232

Garfield E (1998) The impact factor and using it correctly. Unfallchirurg 101:413–414

Glänzel W (2002) Journal impact measures in bibliometric research. Scientometrics 53(2):171–193

Hassan S, Haddawy P (2013) Measuring international knowledge flows and scholarly impact of scientific research. Scientometrics 97:163–179

Hengl T, Minasny B, Gould M (2009) A geostatistical analysis of geostatistics. Scientometrics 80(2):491–514

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci U S A 102:16569–16572

Leidner JL (2007) Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names. Dissertation.com, ISBN: 978–1581123845., pp 289

Leydesdorff L, Bornmann L (2012) Mapping (USPTO) patent data using overlays to Google maps. J Am Soc Inf Sci Technol 63(7):1442–1458

Leydesdorff L, Persson O (2010) Mapping the geography of science: distribution patterns and networks of relations among cities and institutes. J Am Soc Inf Sci Technol 61:1622–1634

Liu L, Wang XY, Li Y (2008) Research on the digital library construction based on GIS. J Inf 1:133–135

Ma D, Meng L (2009) The change of the number of troops in the Space Time Cube for the Red Army Long March. Proceedings of International Symposium on Spatial Analysis, Spatial-Temporal Data Modeling, and Data Mining (ISSA 2009) (pp. 74720X, 1–10). Wuhan

Ma TC, Cao MK, Wang GF (2011) Analysis of the regional competitiveness of basic research in China based on the National Natural Science Fundation of China. Chin Sci Bull (Chinese Ver) 56:3115–3121

Olechnicka A, Ploszaj A (2010) Mapping the regional science performance evidence from Poland. Collnet J Scientometrics Inf Manag 4:21–27

Pouliquen B, Kimler M, Steinberger R, et al. (2006) Geocoding multilingual texts: Recognition, disambiguation and visualisation. Proceedings of LREC-2006, pp, 53–58

Pritchard A (1969) Statistical bibliography or bibliometrics? J Doc 25: 348–349

Schreiber M (2008) A modification of the h-index: the hm-index accounts for multi-authored manuscripts. J Informetrics 2:211–216

Smith T, Andresen D, Carver L, Dolin R, Fischer C, Frew J, Goodchild M, Ibarra O, Kemp R, Kothuri R, Larsgaard M, Manjunath B, Nebert D, Simpson J, Wells A, Yang T, Zheng Q (1996) A digital library for geographically referenced materials. Comput (IEEE) 29(5):54–60

Van Eck NJ, Waltman L (2010) Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84:523–538

Vanclay JK (2012) Impact factor: outdated artefact or stepping-stone to journal certification. Scientometrics 92:211–238

Vinkler P (2010) Indicators are the essence of scientometrics and bibliometrics comments to the book entitled "bibliometrics and citation analysis, from the science citation index to cybermetrics from Nicola De bellis". Scientometrics 85:861–866

Wang XM, Ma MG (2009) Spatial information mining and visualization for Qinghai-Tibet Plateau's literature based on GIS. Proceedings of International Symposium on Spatial Analysis, Spatial-Temporal Data Modeling, and Data Mining (ISSA 2009) (pp. 74920T, 1–8). Wuhan

Wang XM, Li X, Ma MG, Zhang ZQ (2012) Spatial analysis on the geographical information of the scientific literature for qinghai-tibet plateau. Adv Earth Sci 27:1288–1294

Woodruff AG, Paunt C (1994) GIPSY: automatic geographical indexing of documents. J Am Soc Inf Sci Technol 45:645–655

Wu J (2013) Geographical knowledge diffusion and spatial diversity citation rank. Scientometrics 94:181–201

Zhao RY, Xu LM (2010) The knowledge map of the evolution and research frontiers of the bibliometrics. J Libr Sci China 36:60–68