Physica A 466 (2017) 462–475

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Analysis of co-occurrence toponyms in web pages based on complex networks

Xiang Zhong, Jiajun Liu, Yong Gao *, Lun Wu

Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing 100871, China

HIGHLIGHTS

- Proposed the toponym co-occurrence network model.
- Found characteristics of this toponym co-occurrence network, including small world, scale-free and disassortative.
- Presented two new methods to extract core toponyms from web pages.

ARTICLE INFO

Article history:
Received 7 January 2016
Received in revised form 8 July 2016
Available online 3 October 2016

Keywords:
Toponym
Toponym co-occurrence
Complex network
Link analysis
Geographic information retrieval

ABSTRACT

A large number of geographical toponyms exist in web pages and other documents, providing abundant geographical resources for GIS. It is very common for toponyms to co-occur in the same documents. To investigate these relations associated with geographic entities, a novel complex network model for co-occurrence toponyms is proposed. Then, 12 toponym co-occurrence networks are constructed from the toponym sets extracted from the People's Daily Paper documents of 2010. It is found that two toponyms have a high co-occurrence probability if they are at the same administrative level or if they possess a part-whole relationship. By applying complex network analysis methods to toponym co-occurrence networks, we find the following characteristics. (1) The navigation vertices of the co-occurrence networks can be found by degree centrality analysis. (2) The networks express strong cluster characteristics, and it takes only several steps to reach one vertex from another one, implying that the networks are small-world graphs. (3) The degree distribution satisfies the power law with an exponent of 1.7, so the networks are free-scale. (4) The networks are disassortative and have similar assortative modes, with assortative exponents of approximately 0.18 and assortative indexes less than 0. (5) The frequency of toponym co-occurrence is weakly negatively correlated with geographic distance, but more strongly negatively correlated with administrative hierarchical distance. Considering the toponym frequencies and co-occurrence relationships, a novel method based on link analysis is presented to extract the core toponyms from web pages. This method is suitable and effective for geographical information retrieval.

1. Introduction

The Internet has become an important tool by which to deliver and exchange information, undergoing rapid development and attaining high popularity, and contains large amounts of geospatially referenced information. Location information is often associated with these web pages, such as news articles and blog posts [1]. Meanwhile, individuals can contribute

* Corresponding author.
E-mail address: gaoyong@pku.edu.cn (Y. Gao).

http://dx.doi.org/10.1016/j.physa.2016.09.024
0378-4371/© 2016 Elsevier B.V. All rights reserved.

geographical information to the web by sharing their positions, which brings enormous opportunities as well as challenges for the deep mining and knowledge discovery of geographical information [2–4].

Toponyms are among the most common types of geographical information in web documents [5]. The term refers to a proprietary name for a geographic entity with a specific orientation and a special geographical area. It can also be defined as a description of a location, which is a predetermined place [6]. Toponyms can generally be found in newspapers, travel notes, and so on. If two toponyms appear in the same web page, they are co-occurring. Co-occurring toponyms have strong relevance (such as political, economic and spatial relation) when they appear in the same topic of web texts. The more frequently one pair of toponyms co-occurs, the stronger the relevance between them. Co-occurring toponyms are first applied to dissolve the ambiguity of geographical names, namely, in toponym disambiguation [7–10], and are also used to research geographical relatedness. Liu et al. present a method to capture the relatedness between geographical entities based on their occurrences on web pages [11]. Analyzing the co-occurrence and topological distance between all pairs of geographical entities indicates that spatially close toponyms generally have similar co-occurrence patterns, and the frequency of co-occurrence exhibits a distance decay effect under the power law.

The co-occurrence phenomenon was first proposed in natural language processing to study word distribution. Word co-occurrence is defined as several words appearing together with a certain frequency in the same text document. Co-occurrence is widely applied in linguistics, scientific research cooperation and other fields, and has made laudable achievements. Brian Roark et al. present an algorithm for extracting potential entries for a category from an on-line corpus, based on a small set of exemplars [12]. Liang et al. construct a Chinese character (word) co-occurrence network based on the usage features of Chinese characters in poems and analyze the overall structure characteristics of poem co-occurrence networks from a linguistic perspective [13–15]. White and Griffith investigate the cooperation relationships of co-authors based on the number of papers the authors cite together, and they offer a new technique that contributes to the understanding of intellectual structure in science and possibly in other areas [16]. Leydesdorff et al. present both a global map with the functionality of a Google Map (e.g., zooming) and network maps based on normalized relations [17], and Qiu et al. provide a better comprehension of author interaction and contribute to the cognitive application of author co-occurrence network analysis [18]. Henry Small proposes a method to measure the relationship between two co-citation documents and provides a new approach to the study of SDI profiles [19]. Moreover, Rada Mihalcea extracts the key words and sentences from documents based on considering the frequency of co-words, which could automatically create an index for document collection [20].

In addition to many toponyms co-occurring in the same document, a toponym may also appear in many documents simultaneously. As a result, co-occurring toponyms are connected by a document, and documents are associated when they share the same toponym. The iterative associations will further lead to indirect connections between toponyms and finally form a network structure. Therefore, a novel complex network model is proposed in this paper to model the co-occurrence of toponyms. Twelve sample co-occurrence networks of Chinese geographical names are constructed and analyzed to investigate their structural characteristics, including centrality, degree distributions, the small-world feature and assortativeness. Furthermore, a link-based method is presented to find core toponyms, which can help to extract rich geographical information from massive web documents.

This paper is organized as follows. Section 2 proposes the co-occurrence toponym network model, using the frequency and the co-occurrence relationship of geographic names. Section 3 constructs 12 sample co-occurrence networks and analyzes their structural features. Section 4 gives an effective link-based method to obtain the core toponyms from the web page collection, and experiments are conducted to verify this technical solution. The discussions and conclusions are drawn in Sections 5 and 6.

2. Modeling toponym co-occurrence networks

Thousands of web pages form the major sources in the Internet. Web pages are organized by a document collection, which are represented as

$$W = \{D_1, D_2, \ldots, D_n\}. \tag{1}$$

A web page is a collection of toponyms, represented as

$$D_i = \{x_1, x_2, \ldots, x_n\} \tag{2}$$

where a toponym is represented as x_i.

The appearance of two toponyms in the same document is defined as toponym co-occurrence. Thus, every two toponyms in a document, i.e., $\forall x_p, x_q \in D_i$, are co-occurring. A toponym participating in co-occurrence might also exist in many other documents, so these documents could be connected indirectly by sharing toponyms. A strong interrelated correlation between co-occurrence toponyms shows when they appear together in web pages. Considering the co-occurrence relationship and the transmission effect of toponyms, a graph structure is constructed, namely a toponym co-occurrence network. In this network, the toponyms extracted from web pages form the vertices of the graph, and their co-occurrence relationships are expressed by the edges. Two toponyms are linked by an edge if they co-occur in the same web page.

Formally, let $G = (V, E)$ be an undirected graph in which the out-degree of a vertex equals the in-degree of the vertex. The defined graph contains a set of vertices V and a set of edges E, where E is a subset of $V \times V$. The number of vertices N is