Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

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

Semantic overlay network for large-scale spatial information indexing

Zhiqiang Zou [a,b,c], Yue Wang [a], Kai Cao [d,e,*,1], Tianshan Qu [a], Zhongmin Wang [a]

[a] College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, China
[b] Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, Jiangsu 210003, China
[c] Key Lab of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Ministry of Education Jiangsu Province, Nanjing, Jiangsu 210003, China
[d] World History Center, University of Pittsburgh, Pittsburgh, 230 South Bouquet Street, Pittsburgh, PA 15260, USA
[e] Center for Geographic Analysis, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 6 September 2012
Received in revised form
3 April 2013
Accepted 22 April 2013
Available online 27 April 2013

Keywords:
Semantic similarity
Overlay networks
Spatial information indexing
Self-organizing algorithms
Distributed systems

ABSTRACT

The increased demand for online services of spatial information poses new challenges to the combined filed of Computer Science and Geographic Information Science. Amongst others, these include fast indexing of spatial data in distributed networks. In this paper we propose a novel semantic overlay network for large-scale multi-dimensional spatial information indexing, called SON_LSII, which has a hybrid structure integrating a semantic quad-tree and Chord ring. The SON_LSII is a small world overlay network that achieves a very competitive trade-off between indexing efficiency and maintenance overhead. To create SON_LSII, we use an effective semantic clustering strategy that considers two aspects, i.e., the semantic of spatial information that peer holds in overlay network and physical network performances. Based on SON_LSII, a mapping method is used to reduce the multi-dimensional features into a single dimension and an efficient indexing algorithm is presented to support complex range queries of the spatial information with a massive number of concurrent users. The results from extensive experiments demonstrate that SON_LSII is superior to existing overlay networks in various respects, including scalability, maintenance, rate of indexing hits, indexing logical hops, and adaptability. Thus, the proposed SON_LSII can be used for large-scale spatial information indexing.

Published by Elsevier Ltd.

1. Introduction

1.1. Background and related work

Overlay networks are recognized as a way of improving Internet services in many fields, most of which only related to one-dimensional data (Hu et al., 2008; Tanin et al., 2007; Jelasity et al., 2009). With the increase in demand for online services of spatial information, the challenges of fast indexing in distributed networks has become a hot research topic in the combined field of Computer Science and Geographic Information Science. The indexing information involves multi-dimensional data, the corresponding online services of which are becoming increasingly more complicated, especially in that they require complex spatial extent data (e.g. BBox, Bounding Box) with a massive number of concurrent users through network browser. For example, there exist a massive number of concurrent queries for spatial extent data about the SARS (Severe Acute Respiratory Syndrome) in the spring of 2003. The GIS (Geographic Information Systems) based on browser/server (Gogu et al., 2006) cannot efficiently support these large-scale spatial information queries and indexing since the central GIS server unavoidably suffers with hotspot bottleneck and single point of failure (Wu et al., 2010). Here term large-scale means the number of concurrent users is very large, maybe up to thousands or even millions. There exists a contradiction between hotspot bottleneck of GIS server and a massive number of concurrent users for spatial information queries. This scientific problem has not yet been solved completely. Therefore, various new methods have been investigated to address this problem.

Distributed computing technologies such as Grid computing, Cloud computing and overlay networks are mainly used to support large-scale spatial information indexing. In Grid computing, there is often a central management for resource management and work load allocation. Therefore, compared with Grid computing, overlay networks have the advantages of dynamicity and scalable self-configuration (Foster and Iamnitchi, 2003). Cloud computing is another computing model that distributes computing tasks in a resource pool based on a central computer cluster (Mateescu et al., 2011). Contrary to Cloud computing, overlay networks can make full use of idle computing resources of many endpoints (Tanin et al., 2007; Jelasity et al., 2005, 2009; Kantere et al., 2009; Wu et al., 2010; Zou et al., 2011).

Regarding spatial data network processing, Tanin et al. (2006, 2007) presented their preliminary work, DQTChord, which is the work most closely related to our study. In this paper, we compare

[*] Corresponding author.
E-mail addresses: kaicao@pitt.edu, kaicao@fas.harvard.edu (K. Cao).
[1] Postdoctoral Research Associate at University of Pittsburgh and Visiting Research Fellow at Harvard University.

0098-3004/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.cageo.2013.04.019

our approach with DQTChord, in which the spatial data is first divided based on Distributed Quad-Tree, and then mapped onto a Chord. Our approach differs from theirs in that we use a hybrid structure for the overlay network, which includes two layers and considers semantic information. Peers with similar semantics in our semantic overlay network are organized into same peer clusters, which make sharing information among peers more efficiently. The benefits of our approach include a decreased indexing load and increased indexing efficiency. Other related works include (Hu et al., 2008), which uses the P2P framework FLoD to 3D scene streaming for virtual globe applications, and Kantere et al., 2009, which use SPATIALP2P, a totally decentralized framework for spatial data built on top of a one-dimensional DHT. In (Zhang et al., 2009) a swift tree structure for multidimensional data indexing is proposed, which has a query efficiency of $O(logN)$ in terms of routing hops and an low maintenance cost. Our spatial information indexing is substantially different. It is based on a semantic overlay network with a multi-layer structure.

As for semantic overlays (Resnik, 1999; Doulkeridis et al., 2007; Tirado et al., 2010), Resnik (1999) presented a measure based on the notion of information contexts. An algorithm for the distributed and decentralized construction of hierarchical SONs in unstructured P2P networks is presented by Doulkeridis et al. (2007), which is a distributed method of clustering content in a recursive way. Tirado et al. (2010) proposed an affinity AP2P, which is similar to our approach. It uses a novel affinity-based metric to estimate the distance between clusters of nodes sharing similar content. However, there are two main differences from AP2P. First, we define the different semantic similarity function for spatial information. Second, we introduce the epidemic protocol for overlay construction.

There is a large amount of general purpose literature on epidemic protocols and small world networks. Some authors present general protocols based on gossip and the theoretic foundation of epidemics while our work focuses on spatial information indexing (Jelasity et al., 2005, 2009; Demers et al., 1987; Fabrikant et al., 2002; Renesse et al., 2008). Other authors introduce the theory and methods for small world networks, which is the theoretical foundation for our research work (Watts and Strogatz, 1998; Li et al., 2008).

1.2. Main contributions

In this paper, we introduce a spatial data indexing mechanism, which is based on a semantic overlay network. This network is connected by logical links between peers in a semantic quad-tree topology, which is a small world network with a small average path length and a large clustering coefficient. In our semantic quad-tree, peers with similar semantics are organized into same peer clusters, which can be further adaptively arranged into a small world network. According to this semantic proximity, we can dynamically construct the topology of an overlay network in a robust, dependable manner based on our previous work, i.e., HPSIN and SDI-CDQT (Wu et al., 2010; Zou et al., 2011). In (Wu et al., 2010) a HPSIN was proposed by us, which combines distributed quad-tree with distributed hash table (DHT) to maintain both query efficiency and system load balance. Based on our HPSIN, we further propose a SDI-CDQT extending the quad-tree layer in HPSIN by semantic clustering strategy (Zou et al., 2011), which makes the peers with similar interesting belong to the same cluster. However, there exist insufficiencies under a massive number of concurrent users indexing since both HPSIN and SDI-CDQT did not effectively consider the construction of overlay network for large-scale spatial information indexing. To our best knowledge, there lacks effective fast construction protocol of overlay network and the corresponding theoretical analysis for spatial information indexing.

The main contributions of this paper are thereby four-fold

1) An effective semantic clustering strategy based on a uniform semantic metric is adopted that make peers with similar interests organized in same clusters.
2) A new semantic overlay network and the corresponding construction protocol based on gossip for large-scale spatial information indexing (SON_LSII) are proposed. Furthermore, the theoretical analysis of the convergence speed of SON_LSII is formally presented from four cases. The SON_LSII supports a massive number of concurrent users indexing and is a small world overlay network that achieves a very competitive trade-off between indexing efficiency and maintenance overhead.
3) We propose efficient algorithms to support large-scale range queries of the spatial information in a parallel manner based on SON_LSII framework.
4) Extensive experiments are conducted to evaluate the performance of SON_LSII with regard to various aspects, including scalability, rate of indexing hits, indexing logical hops, adaptability, and so on.

1.3. Paper structure

The rest of this paper is organized as follows: Section 2 presents an overview of our system. In Section 3, we discuss our theoretical model of a semantic overlay network and an analysis thereof. Section 4 details the large-scale spatial information indexing algorithm based on a semantic overlay network. In Section 5 we give the theoretical and experimental results that characterize key properties of this overlay network and a performance analysis of our algorithm. Section 6 presents our conclusions and future work.

2. System overview

Before describing SON_LSII model in detail, we present an overview of our system. While there exist some different points from HPSIN as described in Section 1.3, the system structure of SON_LSII is similar to HPSIN (Wu et al., 2010), which is a hybrid structure including Chord Ring Layer and Semantic Quad-tree Layer as depicted in Fig. 1. Differing from the DQTChord (Distributed Quad-Tree Chord) structure (Tanin et al., 2007), our hybrid structure stores subspace data directly in a semantic quad-tree rather than a Chord ring to reduce the communication cost and enhance indexing efficiency.

Fig. 1 depicts the hybrid structure of SON_LSII, which is organized in layered clustering method. First, we place the head peer of the semantic cluster into the Chord ring layer according to the semantics of spatial information that peer holds. And then, in semantic quad-tree, we let each head peer maintain a cluster of four neighboring peers based on a uniform k-element Semantic Vector (***SV***) where k is some constant. The two main aspects of the semantic space correspond to two sub-***SVs***, $\boldsymbol{SV_1}$ and $\boldsymbol{SV_2}$. $\boldsymbol{SV}=\{\boldsymbol{SV_1}, \boldsymbol{SV_2}\}=\{\{e_{11}, e_{12}, \ldots, e_{1k_1}\}, \{e_{21}, e_{22}, \ldots, e_{2k_2}\}\}$, $k_1+k_2=k$. Note that Semantic Vector corresponds to a point in the semantic space while the word “semantic” here means that the element in SV and the point in the semantic space both contain some concrete feature or attribute, which is represented by one element e_{ij} in the ***SV*** with the data object including a $weight_{ij}$. $\boldsymbol{SV_1}$ and $\boldsymbol{SV_2}$ correspond to two aspects, respectively, i.e., the semantic of spatial information that peer holds in overlay network and physical network performances, such as online time and bandwidth. For an example, if there are three peers A,B,C, peer A requests some