Computers & Geosciences 106 (2017) 150–163

Contents lists available at ScienceDirect

Computers & Geosciences

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

Research paper

A feature selection approach towards progressive vector transmission over the Internet

Ru Miao[a], Jia Song[b,d,*], Min Feng[c]

[a] School of Computer and Information Engineering, School of Economics postdoctoral research station, Henan University, Kaifeng 475004, China
[b] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Beijing 100101, China
[c] Global Land Cover Facility, Department of Geographical Sciences, University of Maryland 4321 Hartwick Road STE 400, College Park, MD 20740, USA
[d] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

ARTICLE INFO

Keywords:
Progressive transmission
Feature selection
Amount of information
Real-time transport protocol

ABSTRACT

WebGIS has been applied for visualizing and sharing geospatial information popularly over the Internet. In order to improve the efficiency of the client applications, the web-based progressive vector transmission approach is proposed. Important features should be selected and transferred firstly, and the methods for measuring the importance of features should be further considered in the progressive transmission. However, studies on progressive transmission for large-volume vector data have mostly focused on map generalization in the field of cartography, but rarely discussed on the selection of geographic features quantitatively. This paper applies information theory for measuring the feature importance of vector maps. A measurement model for the amount of information of vector features is defined based upon the amount of information for dealing with feature selection issues. The measurement model involves geometry factor, spatial distribution factor and thematic attribute factor. Moreover, a real-time transport protocol (RTP)-based progressive transmission method is then presented to improve the transmission of vector data. To clearly demonstrate the essential methodology and key techniques, a prototype for web-based progressive vector transmission is presented, and an experiment of progressive selection and transmission for vector features is conducted. The experimental results indicate that our approach clearly improves the performance and end-user experience of delivering and manipulating large vector data over the Internet.

1. Introduction

Over the past decade, web-based geographic information systems (WebGIS) have been widely adopted for various applications to visualize and share geospatial information over the Internet. The browser-based WebGIS is simpler than client-based GIS application in terms of its data handling capability, which is an advantage when using various types of devices, including PCs, laptops, and mobile phones. In order to reduce the memory and computing needs for data storage, visualization, and analysis, WebGIS applications are usually closely coupled with servers that perform the heavy processing and storage. Therefore, data exchange between the server and browser is expected to be extensive. Because transferring large datasets will likely lead to longer wait times, optimizing the data transfer to reduce the system response time is a challenging issue for WebGIS.

Tiled maps have been commonly adopted as a solution for visualization in WebGIS (e.g., Google Maps, Microsoft Bing Map, and OpenStreetMap) (Crampton, 2009). Instead of providing real data, the tiled maps present data previews that are prebuilt for tiling extents at selected zoom levels. The tiled maps usually have short response times because they avoid the processing time used to create a map from data. However, tiled maps do not provide real data, and they prohibit applications that require real data, such as cases that involve two-way interactions between the system and user. Compared with tiled maps, vector maps are able to implement data querying, editing and spatial analysis on the client side, and satisfy the needs for personalizing map (Ballatore and Bertolotto, 2015). However, the capacity to handle and on-the-fly mapping large-volume vector data on the client side, particularly on mobile terminals, is limited, and it leads to B/S (Browser/Server)-based vector map being challenging. Therefore, the strategy of progressive transmission has been proposed.

Progressive transmission is first presented by Bertolotto and Egenhofer (1999, 2001): a subset of the data is sent first, and it is then incrementally refined in subsequent stages. Compared with raster

* Corresponding author.
E-mail addresses: mr1015@henu.edu.cn (R. Miao), songj@igsnrr.ac.cn (J. Song), fengm@umd.edu (M. Feng).

http://dx.doi.org/10.1016/j.cageo.2017.06.005
Received 26 March 2016; Received in revised form 3 January 2017; Accepted 5 June 2017
Available online 08 June 2017
0098-3004/

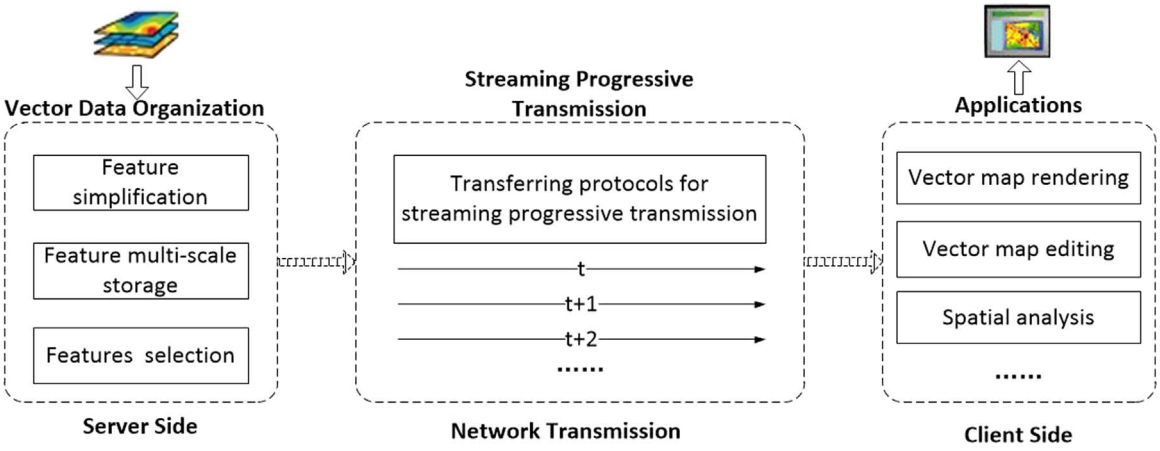


Fig. 1. Framework for web-based progressive vector transmission.

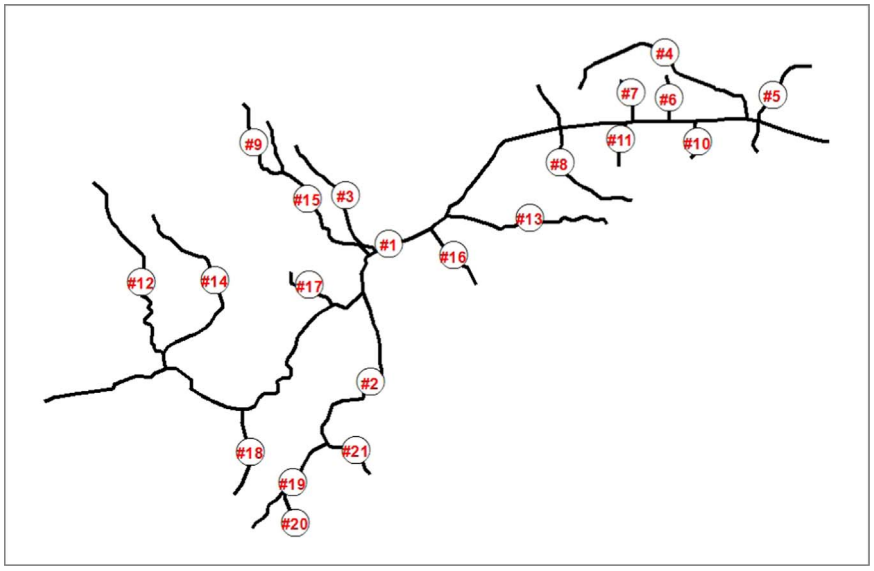


Fig. 2. The map of a river (red labels indicate the serial number of the features).

images, the structure of vector data is more complex (Egenhofer and Franzosa, 1991). Hence, applying progressive transmission to vector data is more complex than applying it to raster data. Buttenfield and McMaster(1991) and Buttenfield 2002) noted that progressive vector transmission should be closely connected with cartographic generalization, including selection, simplification, combination, smoothing and enhancement methods. Han et al. (2003) designed the server/client framework of progressive vector data transmission from technical view. Zhou and Bertolotto (2004) described a sequence of multiple representations generalized by a line simplification algorithm for progressive transmission. Yang et al. (2005) used clustering and multithreading techniques to improve data transmission. Cache and dynamic data management were used to improve client-side interactive performance. Yang, (2004, 2005, 2007a) proposed a modified D-P algorithm to simplify vector data. The modified D-P algorithm maintains a consistent topology of geometric elements on the server side, and restores original vector data on the client side. Ai et al. (2005, 2009) and Ai and Li (2009) suggested a "change accumulation model" to offer an efficient navigation guide with an efficient multiple representations of spatial data. In order to achieve effective transmission of large amounts of vector map, Yang et al. (2007b) built a distributed agricultural information system. They used the improved Douglas-Peucker algorithm and a binary line generalization (BLG) tree to simplify vector data for progressive transmission. Haunert et al. (2009) proposed a method based on topological Generalized Area Partitioning (tGAP) structure for transfering a vector map from a server to a mobile client. Zhang et al. (2011) presented an efficient and robust approach to simplify large geographical maps with frame buffers and *Voronoi* diagrams. Jang et al. (2014) presented a compression method based on a bin space partitioning data structure to transmit large amounts of vector map data. Chen et al. (2014) proposed a coding algorithm to develop the progressive transmission related to the multi-vertex.

In contrast to simplification algorithms, selection approach also plays a very important role in progressive transmission and multi-scale web mapping. When requesting a coarse vector map or large-range vector map over the Internet, those small polygons or lines probably would become very small points or be gathered together. They are less important than other geometries with big size. Similarly, features with more important property information can be transmitted first when delivering large-volume vector data. Therefore, studies on feature selection are necessary towards progressive vector transmission. Jiang and Claramunt (2004) proposed a generalization model for selecting characteristic streets in an urban street network. The model uses graph principles as measures for the selection of important streets. Liu et al. (2010) proposed a stroke-based algorithm for road network selection in map generalization, which considers four types of information: statistical, metric, topological, and thematic. A set of measures were selected to quantify these different types of information at various spatial levels. Following Liu's algorithm, other stroke-based, mesh-

Table 1
The amount of information in geometric size for the river data.

ID	$I_G(P_i)$	ID	$I_G(P_i)$	ID	$I_G(P_i)$	ID	$I_G(P_i)$
1	0.47	7	0.02	13	0.07	19	0.05
2	0.09	8	0.08	14	0.09	20	0.02
3	0.06	9	0.04	15	0.08	21	0.03
4	0.1	10	0.02	16	0.03		
5	0.05	11	0.02	17	0.03		
6	0.02	12	0.11	18	0.04		

Table 2
The amount of information and the contribution to the map in every round.

Round	The number of features	The amount of information	Contribution rate
1	7	1.02	67.1%
2	7	0.34	22.4%
3	7	0.16	10.5%
Total	21	1.52	100%

based or combined stroke-mesh algorithms were proposed for road network selection (Park et al., 2013, Benz et al., 2014, Tian et al., 2014). Ying et al. (2011) and Corcoran et al. (2011a, 2011b, 2012) and Corcoran and Mooney (2011) presented selectivity progressive transmission based on topological consistence of vector data. Regarding point cluster generalization, Yan and Li (2013) also described several quantitative measures of point information that can be used as selection methods, i.e., the number of points for statistical information, the importance of thematic information, the *Voronoi* neighbors for topological information, and the distribution range and relative local density for metric information.

However, studies on selection algorithms for vector features mostly focus on a specific thematic feature, and less discuss how to measure the amount of information in common against vector features. Moreover, progressive transmission is a complicated and systematic process, which undoubtedly involves both the server side and the browser side. The feature selection towards progressive transmission over Real-time Transport Protocol (RTP) is illustrated in the rest of the paper. In Section 2, the concept of the amount of information is introduced, and several factors are presented for measuring the amount of information for vector features, followed by a method of calculating the amount of information for vector features. Then with the selection method, a progressive transmission over RTP is illustrated. Section 3 introduces our prototype system and presents the results and analysis. Section 4 and Section 5 present some discussions and conclusion, including the vision for future research.

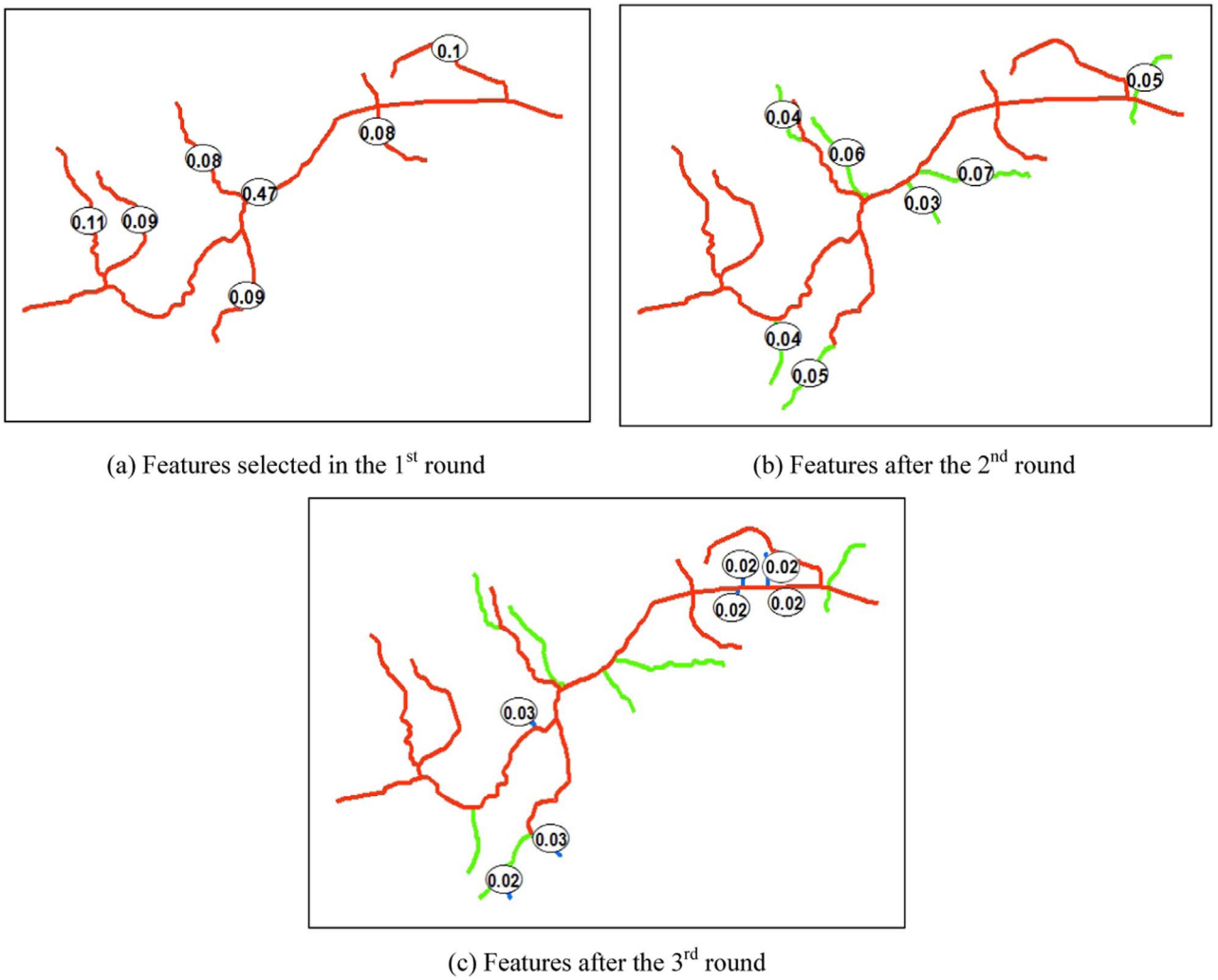
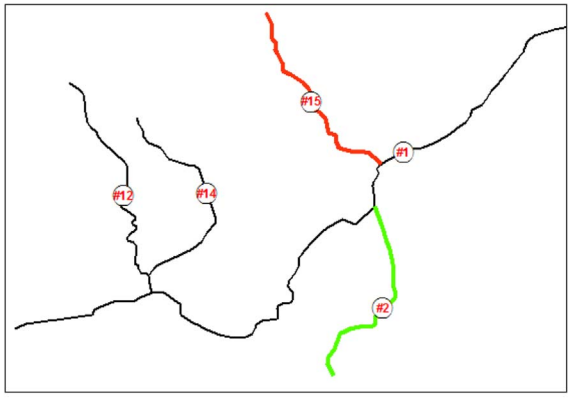
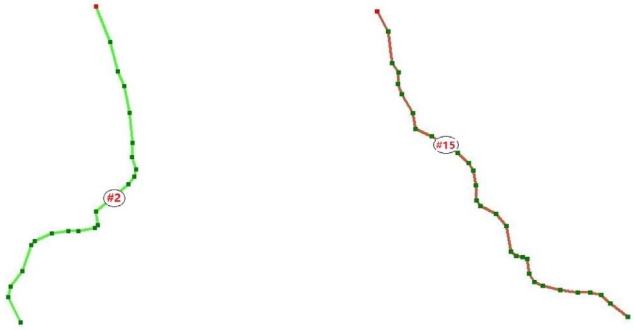


Fig. 3. The results of selection for the river data (the amount of information is labeled on the lines) (a) Features selected in the 1st round (b) Features after the 2nd round (c) Features after the 3rd round.



(a). A portion of the map of the river data



(b). #2 feature of the river data (c). #15 feature of the river data

Fig. 4. The example of geometric complexity for the river data (a). A portion of the map of the river data (b). #2 feature of the river data (c). #15 feature of the river data.

Table 3
The amount of information in geometric complexity.

ID	Length (m)	Count of points	The amount of information	
			Geometric size	Geometric complexity
2	10854.5	22	0.09	0.78
15	10531.3	30	0.08	1.27

2. Methodology

Web-based progressive vector transmission includes data processes at the server side, network transmission and web mapping at the browser side, as shown in Fig. 1. Data processes has two levels. One is simplification, which is coordinate-level, the other is selection, which is feature-level. This study mainly focused on selection process. Unlike simplification process, selection does not involve coordinates removal, and topology relationship does not change before and after selection process. Our study on selection process makes vector features being arranged in terms of the importance of features, and relatively important features are transmitted first. The amount of information is first introduced and used for measuring the importance of vector features in this section, and several factors are presented for calculating the amount of information against vector features. Then, a model, which integrates the proposed factors, is presented to rearrange these features for ensuring relatively important features being transmitted first. Finally, a protocol called Real-time Transport Protocol (RTP) is employed to deliver vector features in progressive transmission.

2.1. The amount of information and information entropy

The amount of information, a component of information theory presented by Shannon (1948), is proposed to represent how much information a random event is. It can be used to measure the degree of uncertainty of a source of information. We use this uncertainty of feature information to suggest the importance of features. In mathematics, the amount of information is the function of probability of a random variable, and is defined as follows:

$$I_r(X_i) = -\log_r P(X_i) \tag{1}$$

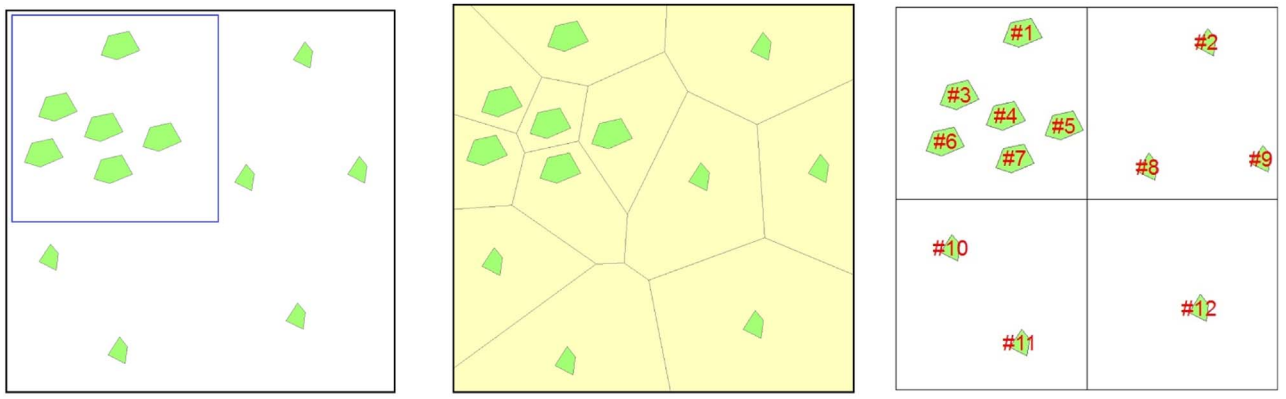
where $P(x_i)(i = 1, 2, 3, ...,n)$ is the probability of a variable of X, and r is the base of the logarithm. "r" is commonly set as 2, Euler's number e, or 10. Generally, r is assigned a value of 2 for ease of processing. In this study, r is set as 2 and the unit of the amount of information is one bit for $r = 2$.

Information entropy is the average amount of information contained in each message received (Stoter et al., 2009). It thus characterizes overall uncertainty regarding the source of information. Shannon explicitly defined the entropy, H (Greek letter eta), of a discrete random variable X with possible values as:

$$H_r(X) = H_r(P_1, P_2, ..., P_n) = -\sum_{i=1}^{n} P(x_i)\log_r P(x_i) \tag{2}$$

where $P(x_i)(i = 1, 2, 3, ...,n)$ is the probability of X with value x_i, $\sum P(x_i) = 1$ *and* $0 \log 0 = 0$.

When applying the amount of information to measure the importance of a feature, considering how to calculate $P(x_i)$ is essential. The calculation model often needs to be suitable for various applications of vector data. Thus, several factors are proposed for the calculation model in the next section.



(a) polygon features (b) polygon features with *Voronoi* diagram (c) polygon features with grid

Fig. 5. Spatial distribution of polygon features (a) polygon features (b) polygon features with *Voronoi* diagram (c) polygon features with grid.

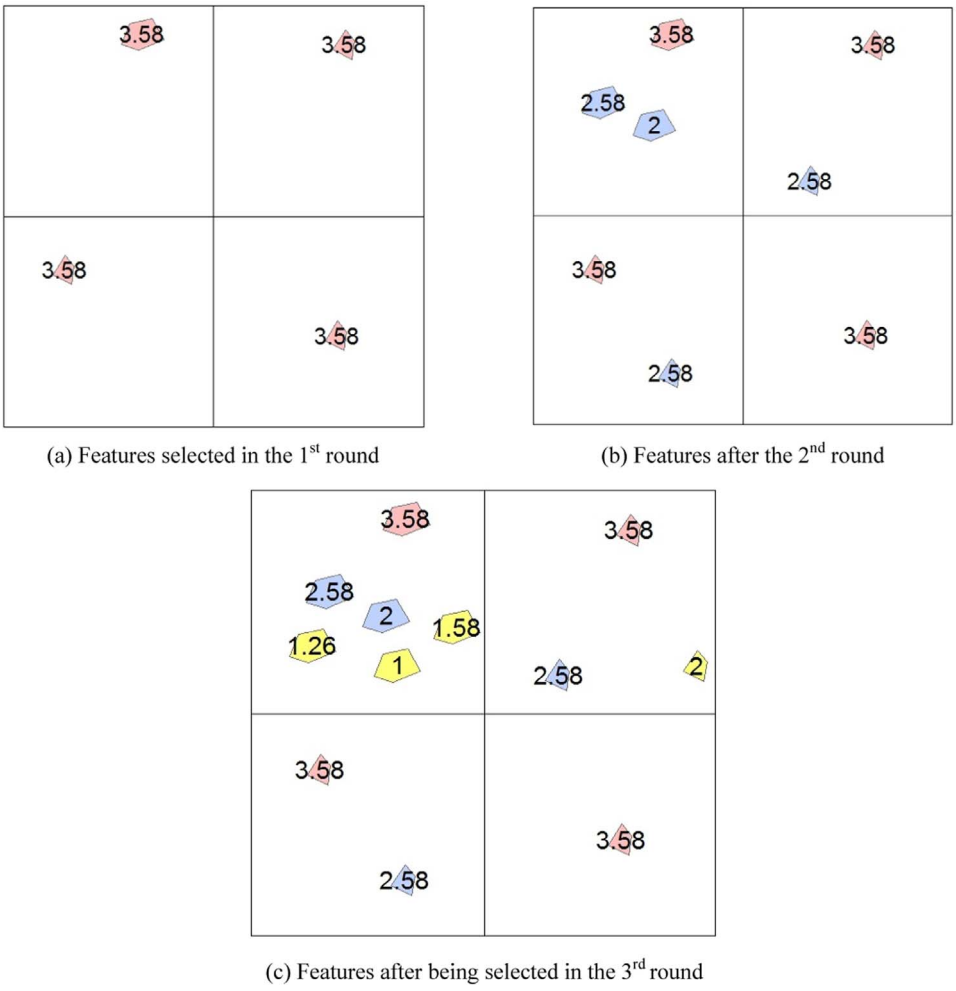


Fig. 6. The selection process of the polygon features (a) Features selected in the 1st round (b) Features after the 2nd round (c) Features after being selected in the 3rd round.

Table 4
The amount of information and contribution rate.

Round	The number of features	The amount of information	Contribution rate
1	4	14.23	47.9%
2	4	9.74	32.6%
3	4	5.84	19.5%
Total	12	29.9	100%

2.2. Measurement factors for feature importance

Considering the amount of information of vector features, features with unusual characteristics in size, attribute, complexity, etc. can be transmitted firstly, and features that have greater contributions to whole structure of vector map can also be transmitted with high priority. Vector data mainly includes geographic spatial information and thematic attribute information. Based on the character of vector data, we present four factors. They are geometric size, geometric complexity, spatial distribution, thematic attribute.

2.2.1. Geometric size factor

The geometric size factor is concerned with the size and shape of feature objects. The length of curves and the area of polygons are of interest. Let N be the count of all feature objects of a map, and G_i be the area of a polygon or the length of a line. Then the amount of geometrical information of the i^{th} feature object of a vector map is defined as:

$$P_i = \frac{G - G_i}{G},\ G = G_1 + G_2 + \ldots + G_N$$
$$I_G(P_i) = -\log P_i = -\log \frac{G - G_i}{G} \quad (3)$$

The above formula reveals that features with larger areas or longer lengths have a higher chance of being selected. It coincides with our assumption that geometric objects with larger areas or longer lengths have a greater effect on mapping and have more information.

Here, an example of a river data is demonstrated for the calculation of the amount of information in geometric size. Fig. 2 shows the map of the original river data with red labels to indicate the serial number of the features. The data has twenty-one line features in total, and the amount of information of the features is listed in Table 1 according to Formula (3).

Let the features being transmitted in three rounds, the results of selection in terms of the amount of information in geometric size are shown in Fig. 3. The amount of information is labeled on the lines in Fig. 3. And the amount of information and the contribution to the map in every round are calculated and listed in Table 2. More than 50% of the amount of information will be selected and transmitted in the first ground, as shown in Table 2.

2.2.2. Geometric complexity factor

In Section 2.2.1, features would have the same amount of information when they have the same length or area. In this situation, we propose another factor called geometric complexity to further distinguish the importance of these features. Polygon features, line features

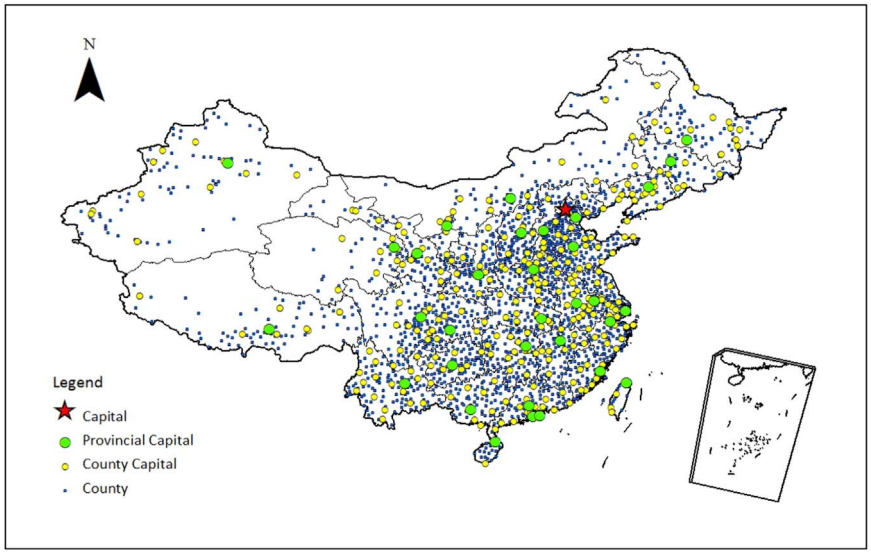


Fig. 7. The map of residential areas of China.

Table 5
The amount of information in thematic attribute.

Residential area level	The number of residential areas	The amount of attribute information	Contribution rate
Capital	1	11.24	54.4%
Provincial capital	33	6.20	30.0%
County capital	297	3.03	14.7%
County	2089	0.21	1.0%
Total	2420	20.68	100%

and point features are essentially composed of a series of points. The number of feature points impacts the geometric complexity. When a feature has the same length or area, more points correspond to greater complexity in shape. Therefore, the average length of a line segment is proposed in this study to indicate the geometric complexity of a feature, and the amount of information in geometric complexity is defined as follows:

$$C_i = \frac{L_i}{N_i}, \quad P_i = \frac{C_i}{C}, C = C_1 + C_2 + \ldots + C_N$$
$$I_C(P_i) = -\log P_i = -\log \frac{C_i}{C} \tag{4}$$

where L_i is the length of the ith line or the boundary length of ith polygon, N_i is the count of feature points. Formula (4) shows that a feature with more points contains a greater amount of information when the length or area of the feature is fixed. Thus, the relatively complex features in shape is able to be selected and transmitted first.

Following the same example in Section 2.2.1, the #2 feature and #15 feature of the river data (Fig. 4(a)) have nearly equal lengths, but they have different counts of points. #2 feature has 22 points, as shown in Fig. 4(b), and #15 feature has 30 points, as shown in Fig. 4(c). They have nearly equal amount of information in geometric size, but #15 feature looks more complex and the amount of information in geometric complexity of #15 feature is greater indeed, as shown in Table 3, which is in accordance with our expectation.

2.2.3. Spatial distribution factor

In addition to the geometric size and complexity, the spatial distribution of features on a map is also important. The more evenly features are distributed, the more information they contain. As shown in Fig. 5, if we are only considering the geometric size factor, the features in the blue box of Fig. 5(a) will be selected firstly because they are bigger in geometric size. However, from the perspective of the overall map, some features outside of the box are smaller in size but may have greater impact on the presentation of the overall map; this viewpoint can be demonstrated by the *Voronoi* diagram

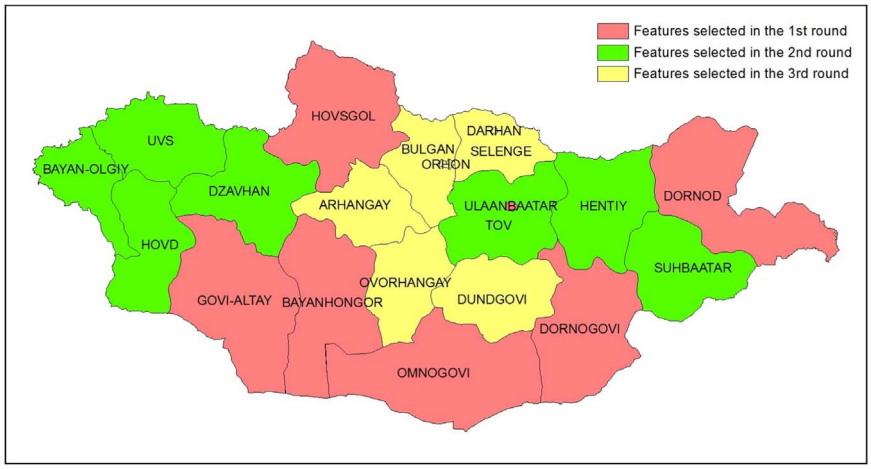


Fig. 8. An example of Provinces of Mongolia.

Table 6
The normalized amount of information for the provinces of Mongolia.

ID	Name	$I_G(P_i)$	$I_{Gnorm}(P_i)$	$I_C(P_i)$	$I_{Cnorm}(P_i)$	$I_S(P_i)$	$I_{Snorm}(P_i)$	$I_A(P_i)$	$I_{Anorm}(P_i)$	$I_{norm}(P_i)$
1	ULAANBAATAR	0.0007	0.0005	3.8431	0.0408	2.3923	0.0298	4.3923	0.7573	0.2071
2	DORNOD	0.1193	0.0801	5.2642	0.0559	4.3923	0.0547	0.0704	0.0121	0.0507
3	GOVI-ALTAY	0.1269	0.0852	4.5813	0.0487	4.3923	0.0547	0.0704	0.0121	0.0502
4	HOVSGOL	0.1088	0.0731	5.1596	0.0548	4.3923	0.0547	0.0704	0.0121	0.0487
5	OMNOGOVI	0.1408	0.0945	4.1585	0.0442	3.3923	0.0422	0.0704	0.0121	0.0483
6	DORNOGOVI	0.1011	0.0679	4.3102	0.0458	4.3923	0.0547	0.0704	0.0121	0.0451
7	BAYANHONGOR	0.1041	0.0699	3.7201	0.0395	4.3923	0.0547	0.0704	0.0121	0.0441
8	HENTIY	0.0829	0.0557	4.5276	0.0481	4.3923	0.0547	0.0704	0.0121	0.0426
9	HOVD	0.07	0.047	4.7135	0.0501	4.3923	0.0547	0.0704	0.0121	0.041
10	TOV	0.0797	0.0535	4.029	0.0428	4.3923	0.0547	0.0704	0.0121	0.0408
11	BAYAN-OLGIY	0.0479	0.0322	5.6361	0.0599	4.3923	0.0547	0.0704	0.0121	0.0397
12	DZAVHAN	0.0853	0.0573	4.437	0.0471	3.3923	0.0422	0.0704	0.0121	0.0397
13	UVS	0.0717	0.0482	5.2633	0.0559	3.3923	0.0422	0.0704	0.0121	0.0396
14	SUHBAATAR	0.0765	0.0514	4.5724	0.0486	3.3923	0.0422	0.0704	0.0121	0.0386
15	DUNDGOVI	0.067	0.045	3.6419	0.0387	4.3923	0.0547	0.0704	0.0121	0.0376
16	ARHANGAY	0.0551	0.037	4.1213	0.0438	4.3923	0.0547	0.0704	0.0121	0.0369
17	SELENGE	0.0432	0.029	4.8299	0.0513	4.3923	0.0547	0.0704	0.0121	0.0368
18	BULGAN	0.0491	0.033	4.5836	0.0487	3.3923	0.0422	0.0704	0.0121	0.034
19	OVORHANGAY	0.0578	0.0388	3.8459	0.0408	3.3923	0.0422	0.0704	0.0121	0.0335
20	DARHAN	0.0003	0.0002	4.5118	0.0479	2.8074	0.0349	0.0704	0.0121	0.0238
21	ORHON	0.0008	0.0006	4.3999	0.0467	2.0704	0.0258	0.0704	0.0121	0.0213
Total		1.49	1.00	94.15	1.00	80.33	1.00	5.80	1.00	1.00

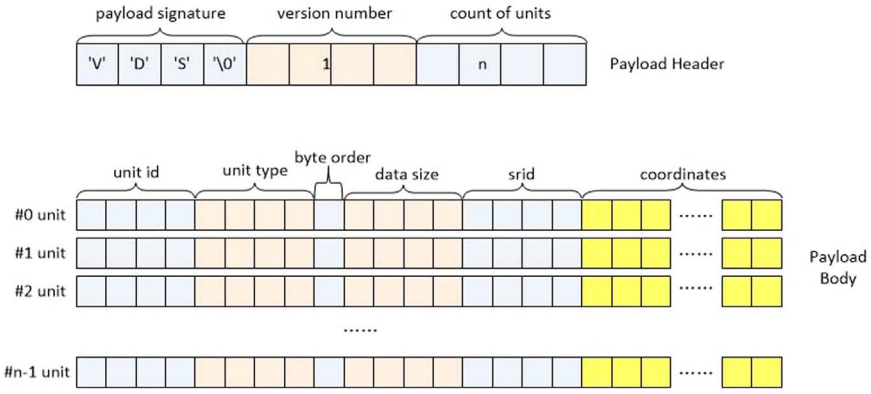


Fig. 9. The payload design for vector features.

Table 7
The payload header for vector features.

Offset	Field	Value	Type
Byte 0	payload signature	'V' (0×56)	UInt8
Byte 1	payload signature	'D' (0×44)	UInt8
Byte 2	payload signature	'S' (0×53)	UInt8
Byte 3	payload signature	'/0′ (0×00)	UInt8
Byte 4	version number	1 (0×01) (Currently)	Int32
Byte 8	the count of body units	The total count of units in the payload body	Int32

Table 8
The payload body for vector features.

Offset	Field	Value	Type
Byte 0	Unit ID	Identification number of a unit, starting from zero.	Int32
Byte 4	Unit Type*	Feature type in data field	Int32
Byte 8	Byte Order	Byte order of geographic features in the data field	UInt8
Byte 9	Data Size	Size of data field in bytes	Int32
Byte 13	SRID	Spatial reference of geographic features	Int32
Byte 17	Data	Features (based on OGC WKB)	Byte[]

* 0: Null shape (metadata), 1: Multi-Point, 2: Multi-Line, 3: Multi-Polygon

(Aurenhammer, 1991). In the Fig. 5(b), the area of *Voronoi* diagram of the features outside of the box are bigger than those inside. Thus, those features with bigger *Voronoi* diagram should be selected firstly in terms of spatial distribution factor.

However, the performance of building *Voronoi* diagrams may not be satisfactory in some scenarios; there are few sophisticated methods to build *Voronoi* diagrams for curves and polygon features. We thus suggest adopting the grid-based approach instead, which involves utilizing a regular spatial grid and calculating the amount of information based on the number of features within every spatial grid cell. Fig. 5(c) shows the result of the polygon features distributed in the grid and the grid partitions. The grid cells are equally divided in spatial area. The number of features inside a cell reflects the density in spatial distribution. In order to avoid features are transmitted centralized in a small area (i.e. a cell), the amount of information in spatial distribution is defined to ensure that features are equally selected and transmitted. Supposing vector features are put inside the grid cell one by one, for the i^{th} feature being put inside the grid cell, P_i is determined by the numbers of features which have already been put inside the same grid cell previously, denoted by S_k. Thus, P_i is represented as follows:

$$P_i = \frac{S_k + 1}{N}$$
$$I_S(P_i) = -\log P_i = -\log \frac{S_k + 1}{N} \quad (5)$$

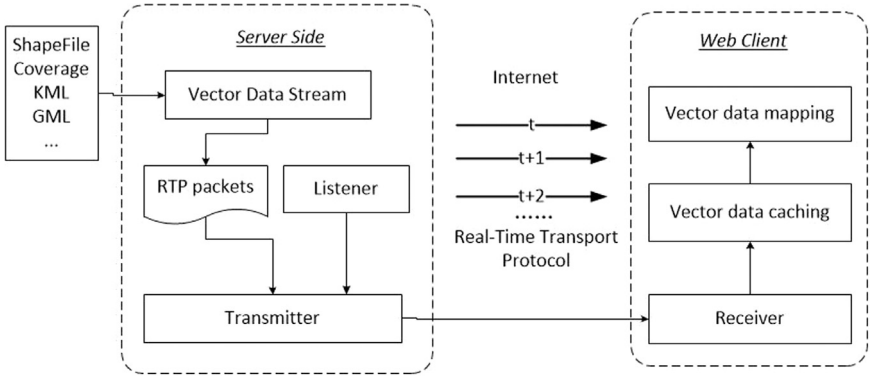


Fig. 10. The prototype system for progressive vector transmission.

where N is the total number of all features. If the ith feature is the first feature falling inside the kth spatial grid cell, S_k equals 0 and P_i is $1/N$. According to Formula (5), the amount of information of the polygon features in Fig. 5 is shown in Fig. 6. Four features are selected in each of rounds, and they are generally selected from every grid cells in each of rounds. The amount of information and contribution rate per round are shown in Table 4.

2.2.4. Thematic attribute factor

Thematic attribute information is another important information of a feature. We analyzed the values of thematic attributes, and find that the attribute values generally has three forms: the first one is enumeration values (e.g.: river level, residential level), the second one is continuous digital values (e.g.: population, temperature), and the third form is free text values (e.g.: address, name). In this study, the thematic attribute factor mainly involves the attributes with enumeration values. An enumeration type defines a range of values, and these values normally have different importance or levels, which are the evidence of which features being transmitted first. Therefore, the amount of thematic attribute information is defined as follows:

$$P_i = \frac{A_i}{A},\ A = A_1 + A_2 + \ldots + A_N$$
$$I_A(P_i) = -\log P_i = -\log \frac{A_i}{A} \tag{6}$$

where P_i depends on the number of features with a certain enumeration attribute. Supposing an enumeration attribute has several enumeration values, A_i (i =1, 2, 3... N) is the number of attribute values corresponding to the i^{th} enumeration value, and A is the total number of attribute values.

Taking a map of residential areas of China as example, which is shown in Fig. 7, residential level is one of attributes, and it has 4 enumeration values. The amount of thematic attribute information of features of the map is shown in Table 5 according to Formula (6).

Table 5 shows that the capital city has the maximum amount of information, followed by provincial capital city, because the number of capital city and provincial capital cities are limited and they are relatively more important.

2.3. Method of measuring the importance of a feature

Given the above measurement factors for the importance of features, it is easy to come up with an integrated measurement model by a linear formula as follows:

$$I(P_i) = a \times I_G(P_i) + b \times I_C(P_i) + c \times I_S(P_i) + d \times I_A(P_i),\ a + b + c + d = 1 \tag{7}$$

where i is the ith feature, and a, b, c, d are the weighted coefficients of these measurement factors. Normally, a, b, c and d can be assigned with the same weights, i.e., the value of a, b, c and d is 1/4.

In terms of the calculation of the amount of information against the above factors, there might be the amount of information in one or some factors is far bigger or smaller than the amount of information in other factors. For example, the amount of information of a feature is 4.5 in geometric size, and it is 0.03 in attribute. If we directly integrate these two results of the amount of information with add operation and same weights "1", the integrated result would be 4.53. Obviously, the amount of information in attribute does not work in this example since it is too small when compared with the amount of information in geometric size. In this situation, the integrated result is not satisfactory. Therefore, we propose a normalization operation to make the amount of information for the above factors scaled in the same range as shown in Formula (8). Then, the Formula (7) is replaced by Formula (9), which is more in line with our goal.

$$I_{norm} = \frac{I(P_i)}{\sum_{k=1}^{N} I_k} \tag{8}$$

$$I_{norm}(P_i) = a \times I_{Gnorm}(P_i) + b \times I_{Cnorm}(P_i) + c \times I_{Snorm}(P_i) + d \times I_{Anorm}(P_i),\ a + b + c + d = 1 \tag{9}$$

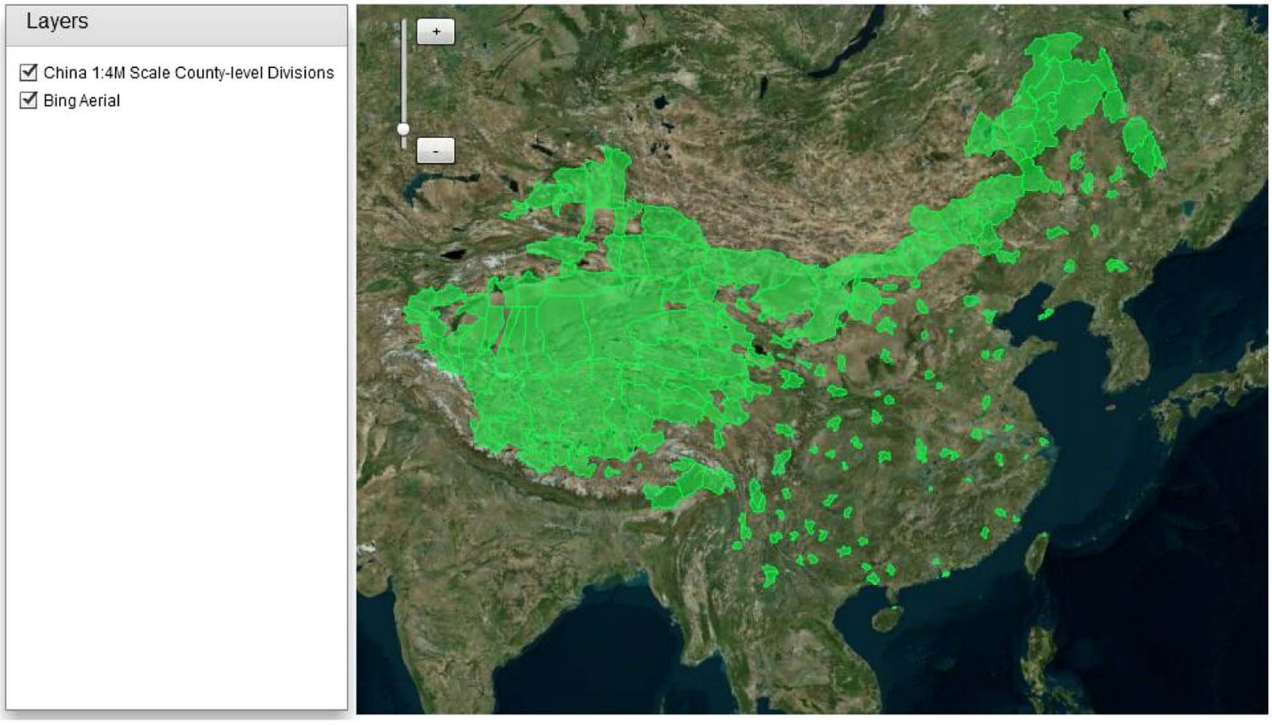
Fig. 8 shows an example of provinces of Mongolia using Formula (9), and Table 6 is the normalized amount of information for this example.

2.4. Progressive transmission based on real-time transport protocol

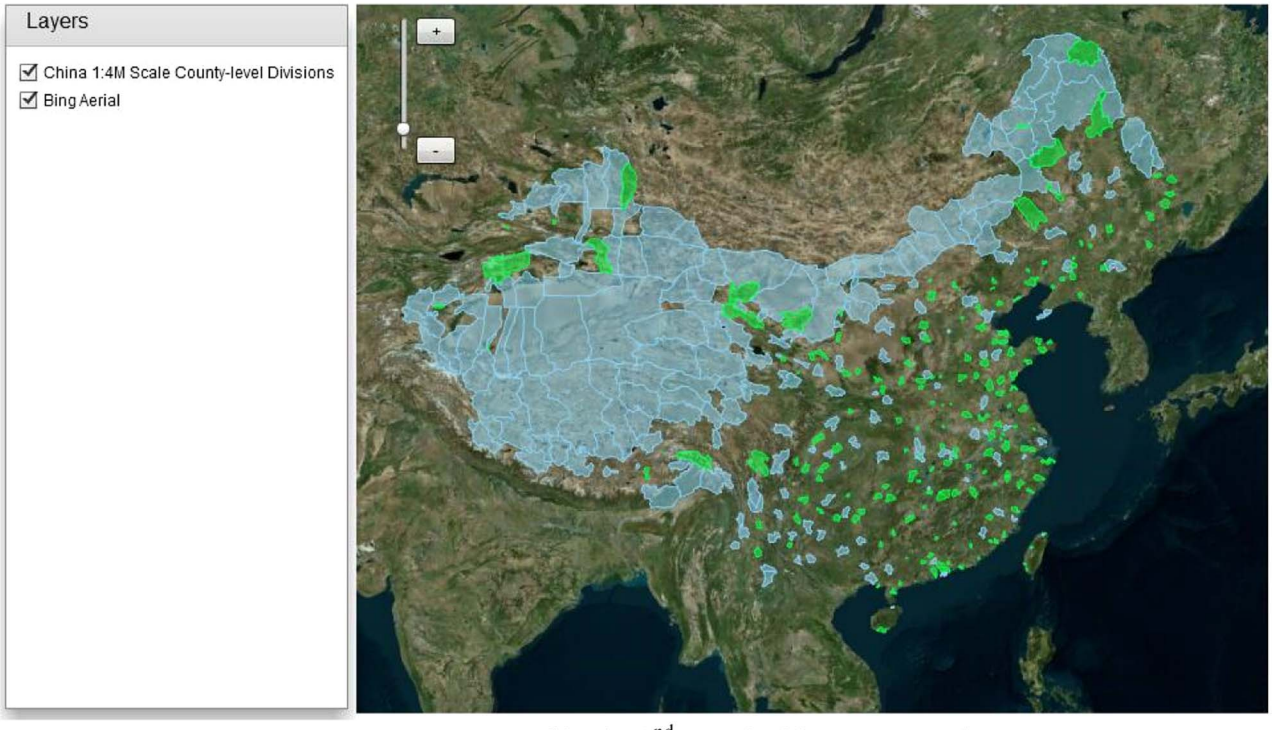
After vector features are marked with the order of transmission based on the measurement of importance of features, progressive transmission is ready to gradually select and deliver features. In this study, Real-time Transport Protocol (RTP), which is originally used to transfer streaming media, is employed in progressive transmission.

A RTP packet payload, which wraps one or more actual transmitted objects, is the key point of the RTP. In progressive vector transmission based on RTP, Vector features are the transmitted objects. A specific RTP packet payload for vector features is designed for achieving delivery round by round. It is composed of a payload header and a payload body, as shown in Fig. 9. The payload header is defined with a payload signature, a version number and the count of units in the payload body, as shown in Table 7. The payload signature is the identifier of vector features; the version number is offered when considering possible changes to the payload in the future. The payload body is composed of a series of body units. The number of body units are recorded in the payload header.

The units of the payload body wrap the actual vector features. Table 8 shows the structure of a body unit. The Unit ID is the identification number of the body unit. The Unit Type is confined to one of values of Metadata, Multi-Point, Multi-Line, and Multi-Polygon. When Unit Type is set with value of Metadata, the data field stores parameter information, such as scale level, spatial range, etc. For other



(a) The 1st round of features to arrive



(b) The 2nd round of features to arrive

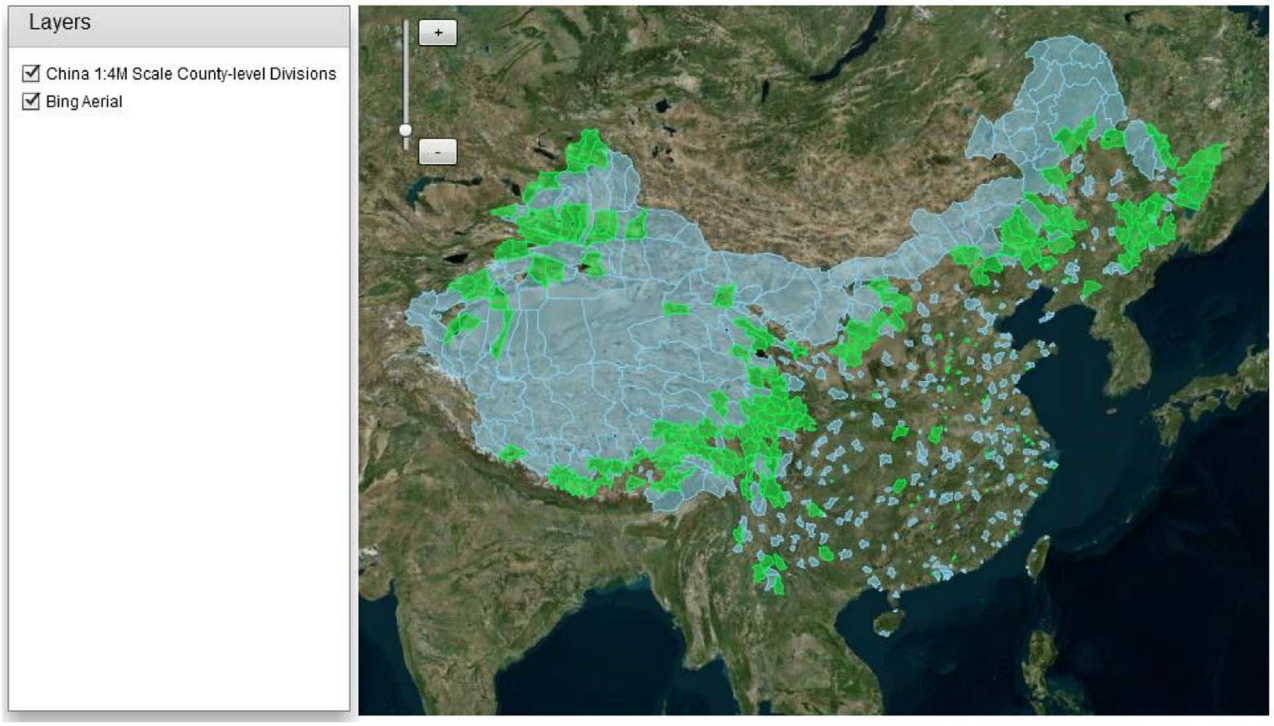
Fig. 11. The progressive selection and web mapping for the China 1:4 M scale county-level divisions (a) The 1st round of features to arrive (b) The 2nd round of features to arrive (c) The 3rd round of features to arrive (d) The 6th round of features to arrive (e) The 9th round of features to arrive (f) The 13th round (the last round) of features to arrive.

geographic features, the data field stores a binary of the vector data itself. We utilize the WKB representation proposed by the Open Geospatial Consortium (OGC) with a few changes; we remove two WKB fields, byteOrder and wkbType, because they are already defined in the Unit Type field and Byte Order field.

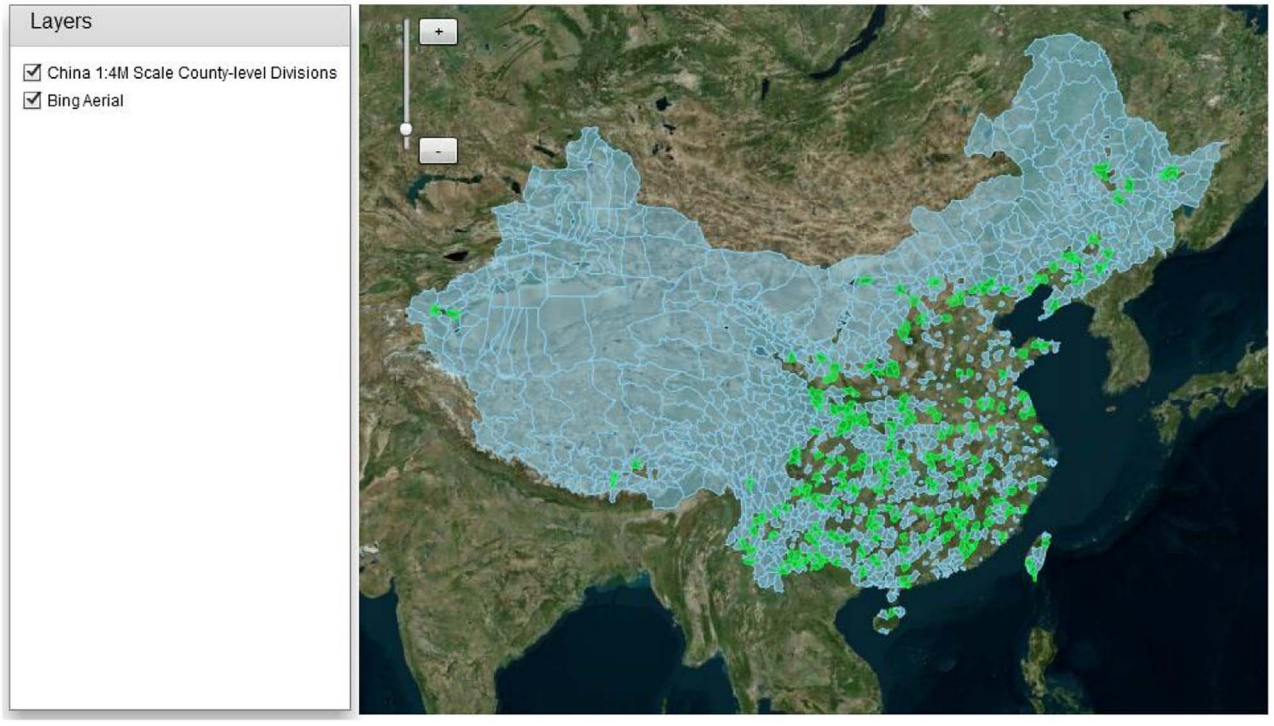
3. Experiment and results

3.1. The prototype of progressive vector transmission

The prototype for web-based progressive vector transmission was developed in C++, C# and web-side script languages. The application is



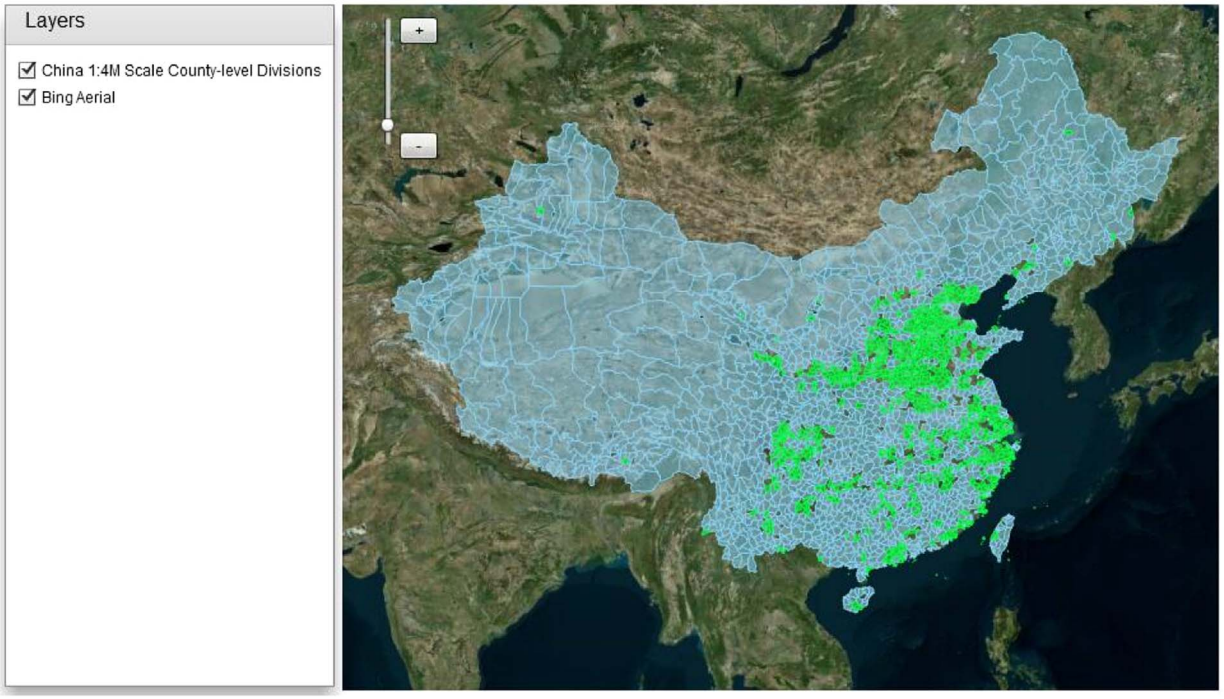
(c) The 3rd round of features to arrive



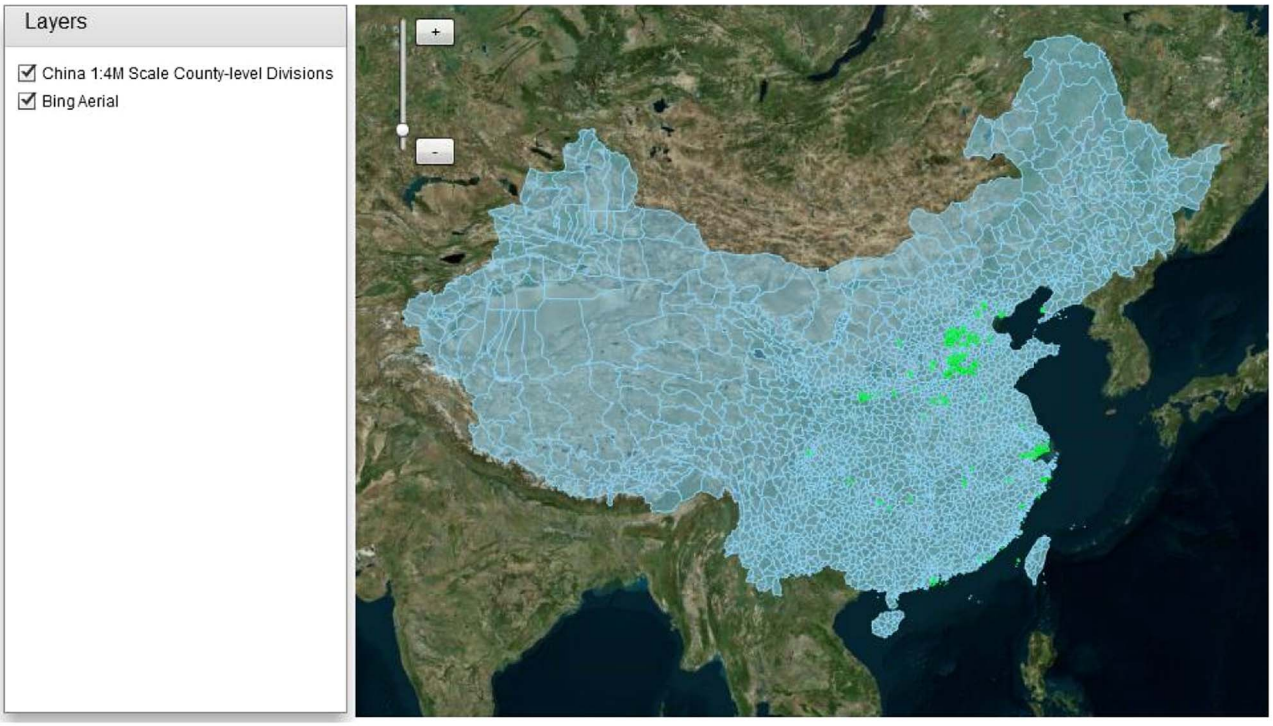
(d) The 6th round of features to arrive

Fig. 11. (*continued*)

based on a B/S (browse/server) architecture. GDAL and JRTPLIB, which are open-source software, are mainly utilized on the server side. GDAL is used for operating geometric entities (Warmerdam, 2008), and JRTPLIB is used for sending packages over the RTP protocol (Liesenborgs, 2007). We also used OpenScales for web-based mapping on the browser side. The prototype system and its components are illustrated in Fig. 10.



(e) The 9th round of features to arrive



(f) The 13th round (the last round) of features to arrive

Fig. 11. (*continued*)

3.2. Experimental results

We choose the data of China 1:4 M scale county-level divisions, which is polygon-type vector data, for the feature selection experiment. The total count of the features is 2525. Fig. 11 shows web mapping results of browser-side. The features of the data are divided into 13 rounds to be delivered in this experiment, and every round transfers 200 features. The blue-color features in the Fig. 11 represent that they have been transferred before this round, and the green-color features represent that they are being transferred in the round. As Fig. 11 shown, features with relatively larger areas are transferred firstly; most counties in Tibet, Xinjiang and Inner Mongolia, China, which have relatively larger areas, were completed after the first rounds of transmissions, and at the same time, some features with high priority

Table 9
The amount of information analysis for the experiment data.

Round	ID	$I_{G.norm}(P_i)$	$I_{C.norm}(P_i)$	$I_{S.norm}(P_i)$	$I_{A.norm}(P_i)$	$I_{norm}(P_i)$
1	1	2.1436	0.0359	0.0545	0.0132	0.5618
	2	1.8306	0.0374	0.0497	0.0132	0.4827
	3	1.4462	0.0363	0.0545	0.0132	0.3876
	...	...				
	200	0.0231	0.0394	0.041	0.1798	0.0708
	total	40.9564	7.6684	9.3881	30.1983	22.0528
2	201	0.0186	0.0377	0.0469	0.1798	0.0708
	202	0.0272	0.0408	0.0352	0.1798	0.0708
	203	0.0282	0.0394	0.0352	0.1798	0.0707
	...	...				
	400	0.0053	0.0394	0.0327	0.1798	0.0643
	total	4.0245	7.9523	7.7847	33.7870	13.3872
3	401	0.1576	0.0366	0.0497	0.0132	0.0643
	402	0.0032	0.0403	0.0337	0.1798	0.0643
	403	0.0012	0.0401	0.0357	0.1798	0.0642
	...	...				
	600	0.0648	0.0373	0.0469	0.0132	0.0405
	total	14.7805	7.7179	8.8593	10.6328	10.4976
.....						
13	2401	0.0041	0.0406	0.0333	0.0132	0.0228
	2402	0.0001	0.0413	0.0367	0.0132	0.0228
	2403	0.0092	0.0408	0.028	0.0132	0.0228
	...	...				
	2525	0.0023	0.0406	0.0268	0.0132	0.0207
	total	0.5091	5.0338	3.8589	1.6482	2.7625
Total		100	100	100	100	100

Table 10
The amount of information analysis for China 1:4 M-scale county-level divisions.

Round No.	Count of transferred features per round	In original order		In sorted order	
		$I_{norm}(P_i)$'	R^k_{trans}'	$I_{norm}(P_i)$	R^k_{trans}
1	200	0.1034	10.3%	0.2205	22.1%
2	200	0.0892	8.9%	0.1339	13.4%
3	200	0.0817	8.2%	0.105	10.5%
4	200	0.0748	7.5%	0.0725	7.3%
5	200	0.0665	6.7%	0.0643	6.4%
6	200	0.0784	7.8%	0.0607	6.1%
7	200	0.0742	7.4%	0.0578	5.8%
8	200	0.0867	8.7%	0.0555	5.6%
9	200	0.0693	6.9%	0.0535	5.3%
10	200	0.069	6.9%	0.0517	5.2%
11	200	0.0796	8.0%	0.0497	4.9%
12	200	0.069	6.9%	0.0474	4.7%
13	125	0.0582	5.8%	0.0275	2.7%
Total	2525	1	100%	1	100%

in attribute factor are first selected at the same time. Additionally, some other counties of China were transferred to ensure a balanced distribution, even though they have relatively small areas.

3.3. Amount of information analysis

We calculated the amount of information of each feature of the experiment data, and listed some results in the Table 9 to demonstrate how the progressive vector selection and transmission are performed. Table 9 shows the sorted order of the 2525 features in terms of the amount of information proposed in this study. $I_{G.norm}(P_i)$,$I_{C.norm}(P_i)$,$I_{S.norm}(P_i)$ and $I_{A.norm}(P_i)$ are the normalized amount of information in geometric size, geometric complex, spatial distribution, and thematic attribute respectively. $I_{norm}(P_i)$ is the total normalized amount of information that integrates $I_{G.norm}(P_i)$, $I_{C.norm}(P_i)$, $I_{S.norm}(P_i)$, $I_{A.norm}(P_i)$ with equal weights "1/4". The features are divided into 13 rounds to deliver, and the number of feature delivered in every round is 200 in this experiment.

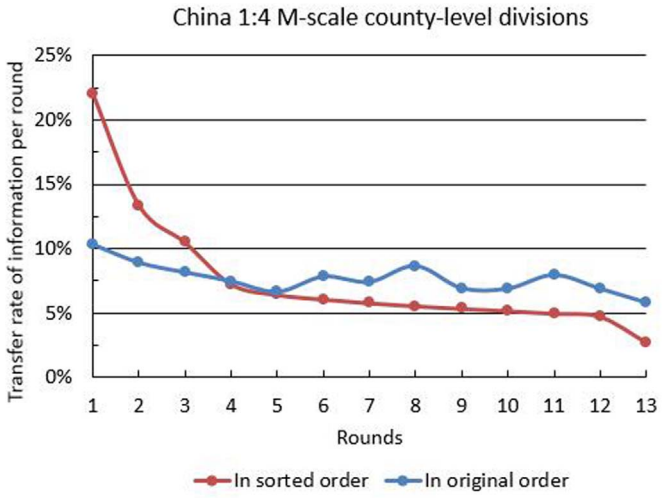


Fig. 12. The comparison of the transfer rate of information for the data.

We further analyze the total amount of information and transfer rate of the amount of information in each of rounds, as shown in Table 10. The transfer rate of the amount of information is written as:

$$R^k_{trans} = \frac{\sum_{i=n\times(k-1)}^{n\times k} I_{norm}(P_i)}{\sum_{i=1}^{N} I_{norm}(P_i)} \times 100\%, \quad k = 1, 2, 3, \ldots; \tag{10}$$

where k represents the k^{th} round, i is the i^{th} feature, n is the count of features per round and N is the total count of all features. In Table 10, normal transmission, in which the features are transferred in original order, and progressive transmission, in which the features are transferred in sorted order, have been compared.

Fig. 12 shows the transfer rate of the amount of information in both original order and sorted order. The red line represents the transfer rate in sorted order. Evidently, the amount of information decreases with the rounds performing, which confirms that the feature selection method based on the amount of information is effective. As the shown in Fig. 12, the amount of information of the first round represents more than 20% of the whole map but only 8% of the number of entire map features. The amount of information being transferred in the first round is also more than other rounds, and the amount of information is getting fewer with the features transmission. That is to say, these features transmitted firstly have more information and reflect the map outlines, and the last features reflect the map details. In other words, most of the map information has finished transfer in previous rounds. At this point, users can perform most of the map analysis operations according to the information on map. Moreover, by comparing the figures regarding the total transfer rate of the amount of information, approximately 50% of the information was transferred in the initial five rounds. In contrast, the transfer rate of the amount of information in original order, which is represented with blue-color line in the Fig. 12, does not varied much with the rounds increasing. Specifically, the amount of information goes a little higher in some rounds, and goes a little lower in other rounds. The phenomenon probably happens to other vector data.

4. Discussion

Vector data is more challenging than raster data in web-based GIS applications since vector data itself is more complex and is not easy to be visualized on the fly, especially when dealing with large amount of vector datasets. In order to improve the performance of web-based GIS application for vector data, the approach of progressive selection and transmission is an effective way, and a feature selection is an essential method towards progressive vector transmission.

Compared with other related studies, we put forward the amount of information of vector features to quantitatively measure the importance of the features. This is a novel approach, which is helpful to select those important features to deliver first in progressive transmission. Moreover, multiple factors (geometric size, geometric complexity, spatial distribution and thematic attribute) are taken into consideration in calculating the amount of information of features, and these factors are proposed under the consideration of general characteristics of vector data. Our calculation method of the amount of information reflects that particular minority features in size, complexity, attribute and spatial distribution have more amount of information. Generally, our method is available for any vector data, and not limited to a particular vector dataset or one kind of data.

Since the amount of information of vector features is an innovative design in this study, we further discuss them as follows: (1) The proposed attribute factor is suitable for those enumeration-type attributes, such as enumerable categories or levels, and it is not suitable for the attributes using continuous values, such as temperature, length, concentration. This qualification is inevitability because attributes have various type of information; (2) *Voronoi* diagram can be used to accurately get the spatial distribution of features. However, considering low efficiency of generating *Voronoi* diagram, we proposed a simple grid-based solution but loss of a little accuracy. The experiment shows that the little loss of accuracy would not impact the overall spatial distribution of features but make the calculation more practical; (3) When calculating the total amount of information, normalization is found indispensable to avoid a certain factor being dominant. Additionally, the weights for the factors are set the same weight in the experiment only for demonstrating how it is performed. Considering the differences of data, the weights can be freely given according to different goals. Taking a road dataset as an example, in order to make roads with all high-level attribute value being selected and transmitted first, much higher weight can be set to the attribute factor.

In addition, progressive vector transmission often involves reciprocating transmission with multiple rounds. How many features are transmitted per round is also depending on different requirements. Features can be delivered in terms of fixed number or data volume for each of rounds, and the number of features delivered per round can also be dynamic. We imagine that the dynamic number of features delivered per round can be determined by the bandwidth of networks.

5. Conclusion

The issue of vector transmission over the Internet and web-based mapping of vector data has presented many challenges. Feature selection is one of essential challenges in progressive transmission. Therefore, we propose a novel feature selection approach, which is based on the measurement for the importance of features quantitatively using the amount of information. The RTP-based progressive transmission is employed in this study to implement progressive selection of features. The experiment results show that the amount of map information transferred in the initial rounds could reach 50% or more of the total, and the first response time is getting very short. End users no longer need to wait for all the vector data to be transferred. The experimental results demonstrate the technical feasibility and usability of this approach.

In the future, we intend to expand our approach to 3-dimension vector data. Additionally, we attempt to research the amount of information of vector layers for determining which layer has priority in progressive vector transmission.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (No. 41201411; No. 41501432); National Special Program on Basic Works for Science and Technology of China (No. 2013FY110900); Fundation of State Key Laboratory of Resources and Environmental Information System (No. O88RA20CYA).

References

Ai T., Ai B., Huang Y.F., 2009. Multi-scale representation of hydrographic network data for progressive transmission over web. In: Proceedings of the 24th International Cartographic Conference.

Ai T., Li J., 2009. Progressive transmission and visualization of vector data over web. In Proceedings of ASPRS Annual Conference. (Cited on page 2).

Ai, T., Li, Z., Liu, Y., 2005. Progressive transmission of vector data based on changes accumulation model. Developments in spatial data handling. Springer Verlag, Berlin-Heidelberg-New York, 85–96.

Aurenhammer, F., 1991. Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput. Surv. 23 (3), 345–405.

Ballatore, A., Bertolotto, M., 2015. Personalizing maps. Commun. ACM 58 (12), 68–74.

Benz, S.A., Weibel, R., 2014. Road network selection for medium scales using an extended stroke-mesh combination algorithm. Cartogr. Geogr. Inf. Sci. 41 (4), 323–339.

Bertolotto, M., Egenhofer, M.J., 2001. Progressive transmission of vector map data over the World Wide Web. GeoInformatica 5 (4), 345–373.

Bertolotto M., Egenhofer M.J., 1999. Progressive vector transmission. In: Proceedings of the 7th ACM international symposium on Advances in geographic information systems. ACM, pp 152–157.

Buttenfield, B.P., 2002. Transmitting Vector Geospatial Data Across the Internet. Geographic Information Science. Springer Verlag, Berlin-Heidelberg-New York, 51–64.

Buttenfield, B.P., McMaster, R.B. (Eds.), 1991. Map Generalization: Making Rules for Knowledge Representation. Longman Scientific & Technical, New York, 150–239.

Chen, M., Wen, Y., Yue, S., 2014. A progressive transmission strategy for GIS vector data under the precondition of pixel losslessness. Arab. J. Geosci. 8 (6), 3461–3475.

Corcoran, P., Mooney, P., Bertolotto, M., et al., 2011a. View-and scale-based progressive transmission of vector data. Comput. Sci. Appl.-ICCSA 2011, 51–62.

Corcoran, P., Mooney, P., Bertolotto, M., 2012. Line Simplification in the Presence of Non-Planar Topological Relationships. Bridging the Geographic Information Sciences. Springer Verlag, Berlin-Heidelberg-New York, 25–42.

Corcoran P., Mooney P., Winstanley A.. et al. 2011b. Effective Vector Data Transmission and Visualization Using HTML5. GIS Research UK (GISRUK), pp 179–183.

Corcoran, P., Mooney, P., 2011c. Topologically consistent selective progressive transmission. Adv. Geoinf. Sci. Chang. World, 519–538.

Crampton, J.W., 2009. Cartography: maps 2.0. Progress. Human. Geogr. 33 (1), 91–100.

Egenhofer, M.J., Franzosa, R.D., 1991. Point-set topological spatial relations. Int. J. Geogr. Inf. Syst. 5 (2), 161–174.

Han H., Tao V., Wu H., 2003. Progressive vector data transmission. In: Proceedings of the 6th AGILE. Lyon, France, pp 103–113.

Haunert, J.H., Dilo, A., van Oosterom, P., 2009. Constrained set-up of the tGAP structure for progressive vector data transfer. Comput. Geosci. 35 (11), 2191–2203.

Jang, B.J., Lee, S.H., Lim, S., Kwon, K.R., 2014. Progressive vector compression for high-accuracy vector map data. Int. J. Geogr. Inf. Sci. 28 (4), 763–779.

Jiang, B., Claramunt, C., 2004. A structural approach to the model generalization of an urban street network. GeoInformatica 8 (2), 157–171.

Liesenborgs J., 2007. JRTPLIB [EB/OL]. ⟨http://research.edm.uhasselt.be/jori/page/CS/Jrtplib.html⟩. Jrtlib, 2006–2016.

Liu, X., Zhan, F., Ai, T., 2010. Road selection based on Voronoi diagrams and "strokes" in map generalization. Int. J. Appl. Earth Obs. Geoinf. 12, S194–S202.

Park, W.J., Lee, Y.M., Yu, K.Y., 2013. The selection methodology of road network data for generalization of digital topographic map. J. Korean Soc. Surv. Geod. Photogramm. Cartogr. 31 (3), 229–238.

Shannon, C.E., 1948. A note on the concept of entropy. Bell Syst. Tech. 27, 379–423.

Stoter, J., Burghardt, D., Duchéne, C., et al., 2009. Methodology for evaluating automated map generalization in commercial software. Comput. Environ. Urban Syst. 33, 311–332.

Tian, J., He, Q., Yan, F., 2014. Formalization and new algorithm of stroke generation in road networks. Geomat. Inf. Sci. Wuhan. Univ. 39 (5), 556–560.
Warmerdam, F., 2008. The geospatial data abstraction library. In: Brent, H., Michael, L.G. (Eds.), Open Source Approaches in Spatial Data Handling. Springer, Berlin, 87–104.
Yan, H., Li, J., 2013. An approach to simplifying point features on maps using the multiplicative weighted Voronoi diagram. J. Spat. Sci. 58 (2), 291–304.
Yang, B.S., Purves, R.S., Weibel, R., 2004. Implementation of progressive transmission algorithms for vector map data in web-based visualization. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., 34.
Yang, B.S., Purves, R., Weibel, R., 2007a. Efficient transmission of vector data over the internet. Int. J. Geogr. Inf. Sci. 21 (2), 215–237.
Yang, B.S., 2005. A multi-resolution model of vector map data for rapid transmission over the Internet. Comput. Geosci. 31 (5), 569–578.
Yang, C., Wong, D.W., Yang, R., et al., 2005. Performance-improving techniques in web-based GIS. Int. J. Geogr. Inf. Sci. 19 (3), 319–342.
Yang, J., Zhang, X., Fan, Y., et al., 2007a. Progressive transmission of vector data in a distributed agricultural information system. N.Z. J. Agric. Res. 50 (5), 1323–1330.
Ying F, Mooney P, Corcoran P, et al., 2011. Selective progressive transmission of vector data. 1–5.
Zhang, L., Zhang, L., Ren, Y., et al., 2011. Transmission and visualization of large geographical maps. Isprs J. Photogramm. Remote Sens. 66 (1), 73–80.
Zhou, M., Bertolotto, M., 2004. A Data Structure for Efficient Transmission of Generalised Vector Maps. Computational Science-ICCS 2004. Springer Verlag, Berlin-Heidelberg-New York, 948–955.