Assessment of MongoDB's Spatial Retrieval Performance

Miaoran Duan [1], Gang Chen [1]*
[1] Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Department of Geographic Information Science, Nanjing University, Nanjing, 210093
*Corresponding author, e-mail: chengang@nju.edu.cn

Abstract—**Based on four sample point sets of 10 thousand (100*100), 250 thousand (500*500), 1 million (1000*1000) and 4 million (2000*2000), we test the space retrieval performance of MongoDB and ArcGIS in different sample quantity condition. The test results show that, after enabling spatial index, spatial retrieve performance of MongoDB is much better than ArcGIS, and the performance gap increases with the increase of sample point set; both of two products have shown strong regularity and stability in all sample point sets under test environment. Finally, experimental results explain the characteristic and application directions of two products.**

Keywords- ArcGIS; MongoDB; NoSQL; space retrieval

I. INTRODUCTION

Relational databases are databases based on relational model. Relational model is composed of relational data structure, relational operation set and relational integrity constraint [1]. Due to the arrival of Web2.0 era, humanized websites have become mainstream and users have not only become an information source, but have been turned from a passive receiver of Internet information into an active creator of Internet information. Because of the high-concurrency and high-load database, massive data storage and access, growing database capacity, and increasingly weakened multi-table correlated subquery, a series of problems insurmountable to relational database have arisen immediately. To solve these problems, NoSQL database came into being at the right moment [2-4].

NoSQL refers to those non-relational and distributed data storage systems that may not follow ACID principles [5]. As one of the representatives of the current NoSQL database products, MongoDB is a kind of documental database based on key-document [6, 7].

With the rapid popularization of various mobile terminals in the recent years, location-based service (LBS) and its related applications have continued increasing in number, and it is also the development trend of society informatization. Domestic and foreign scholars have carried out a lot of researches on the NoSQL (represented by MongoDB) and LBS [8, 9]. Shen Derong et al. summarized the relevant researches of NoSQL database [10]; Wu Fei put forward a concrete thought to solve the technical bottleneck of location storage inquiry by taking MongoDB as a file-level storage [11]; Wang Guanglei optimized the problem about the mal-distribution of the distributed file systems for massive data storage by using consistent Hash Algorithm [12]; Zhang Yao defined the storage model of a geographic information sharing data by using BSON as a descriptive approach to geographic information sharing data [13]. Most of the researches are concentrated on the storage performance and distributed construction of MongoDB, while there aren't many studies on the spatial location indexing which is newly supported by MongoDB.

Based on a quantitative experiment, this paper makes a comparison with the retrieval performance of ArcGIS, the product of ESRI and MongoDB which natively supports spatial location index, to explain the performance characteristics of both in different orders of magnitude of data, and analyze their causes as well as application directions.

II. KEY TECHNICAL ISSUES

ESRI Company's ArcGIS is a mainstream geographic information system software. The Geodatabase under its dominance is based on relational database, and after optimized, it supports rich spatial retrievals and analysis functions, thus solving many problems of relational database in spatial information processing [14, 15]. After Version 2.4, MongoDB, a representative of NoSQL, natively supports spatial location retrieval, and due to its powerful retrieval performance, it yields unusually brilliant results in the LBS application in Web2.0 era. This paper aims to research the advantages and disadvantages of both in spatial location retrieval.

A. ArcGIS spatial retrieval technology

The powerful spatial retrieval and spatial analysis function of ArcGIS almost contain all geographic information service areas, and can create indexes automatically for all spatial data, excellent in both function and performance. ESRI Company has developed its product into a set of universal secondary development component sets – ArcObjects [16]. Retrieval of ArcObjects can realize all functions of ArcGIS. The spatial relational retrievals supported by ArcGIS include seven kinds, which are esriSpatialRelTouches, esriSpatialRelCrosses, esriSpatialRelOverlaps, esriSpatialRelContains, esriSpatialRelContains, esriSpatialRelIntersects and esriSpatialRelRelate. EsriSpatialRelWithin is chosen as a retrieval method for the purpose of corresponding to the spatial retrieval function of MongoDB.

B. MongoDB spatial retrieval technology

MongoDB is known for its excellence in set storage, free mode and high performance, supporting complex query, and supporting complete retrieval. After Version 2.4, it begins to natively support spatial location retrieval and can be directly used for location distance calculation and query. It natively

Supported by the National Natural Science Foundation of China (No. 41271160)

supports geographic location index and can be directly used for location distance calculation and query. It has two commonly-used spatial location indexes.

1) 2d plane coordinate index. It's suitable for plane-based coordinate calculation.

2) 2dsphere geosphere index. It's suitable for spherical geometric operation.

MongoDB takes GeoJSON as a method for spatial location information storage, which is equivalent to a set document. Spatial location retrieval is aimed at this document, and its performance is enhanced substantially after spatial indexing is supported.

III. TEST DATA PREPARATION

A. Test platform

In order to make experimental result objectively showing the actual performance of spatial indexing, a specific console test program is developed for two products respectively. The test is conducted in the same machine and development environment (Tab.1), so that the result should be only driven by the product data driver and influenced by the actual performance.

TABLE I. CONFIGURATION TABLE OF DEVELOPMENT PLATFORM

	MongoDB	**ArcGIS**
Hardware configuration	Xeon E5-2609v2 2.5GHz|16GB	
Operating system	Windows 8.1(64bit)	
SDK	Miscrosoft Visual Studio 2010	
Language	C#	
Tools	MongoDB(native)	ArcObjects
Data driver	mongo-csharp-driver	ArcGIS SDE
Database	MongoDB	Geodatabase

B. Sample dataset design

The objective of the test is to test the spatial retrieval performance of both with various sample sizes. The sample dataset is designed in the evenly distributed point sets in a fixed area, and there are 4 samples in all, including 10 thousand (100*100), 250 thousand (500*500), 1 million (1000*1000) and 4 million (2000*2000) points. The corresponding test program is respectively used to test each sample for 5 times with the degree of coverage from 1% to 100%, and each test is repeated for 5 times, to take a mean value as the result. The test method is to set a rectangular planar zone, and the number of the sample points in the retrieval area respectively corresponds to the Query-class WithinRectangle method in MongoDB as well as the type of esriSpatialRelContains of esriSpatial-RelEnum in ArcGIS, both of which are the most common basic retrieval type.

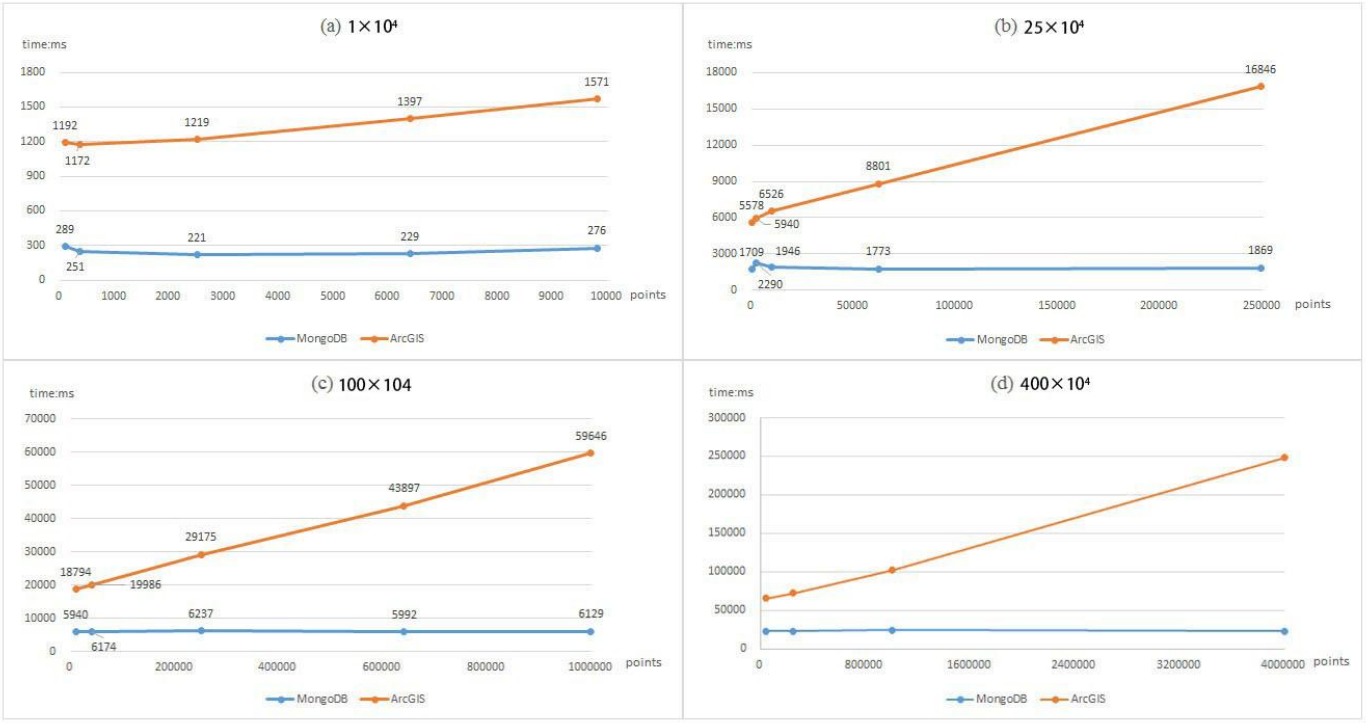

Figure 1. Flow chart of test program

C. *Test flow design*

The test programs runs through 4 phases in all (Fig.1). In order that interference should be suppressed, variable factors are controlled. Only the time consumed by the program in Phase 3 (Data Query) is calculated.

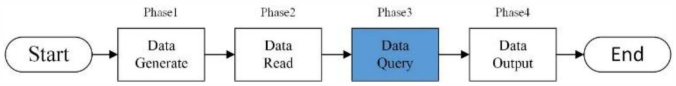

Figure 2. Flow chart of test program

IV. TEST RESULT ANALYSIS

Put the computational results of the test program in Figure 2, and it can be seen: (1) the retrieval efficiency of MongoDB is much higher than ArcGIS in the four orders of magnitude of samples; (2) in each sample, with the increase of retrieved points in number, MongoDB needs constant time, while loner time is spent for ArcGIS.

Continues giving play to its superiority that it has a high retrieval efficiency for documental database, and supports spatial indexing, so its efficiency is very high; secondly, ArcGIS is a huge function set, and the spatial retrieval of point data is merely one of its functions. It realizes function through the inheritance of class, and sets apart many ports for interacting with other classes, so its performance is reduced; finally, about implementation model, ArcGIS identifies point data one by one through SpatialFilter and FeatureCursor, and doesn't support the direct input of retrieval quantity, so its performance is greatly reduced.

In addition, it's seen from Fig.2 that in each order of magnitude of samples, the time needed for ArcGIS increases evenly with the increase of number. Tab.2 is obtained after a trend line is fit for the chart of each sample. In the four samples, the slope always floats around 0.043, and the reliability of R^2 is very high. It can be seen that ArcGIS has a very stable and foreseeable spatial retrieval capability for point features. Regardless of sample sizes, when retrieval quantity increases by 1000, another 43 milliseconds will be consumed.

TABLE II. TREND LINE OF ARCGIS RESULT

Sample	formula	R^2
1w	y = 0.0403x + 1155.6	0.9732
25w	y = 0.0443x + 5871.9	0.9981
100w	y = 0.0409x + 18436	0.9992
400w	y = 0.0467x + 60395	0.9982

The same rule also exists in MongoDB. Since MongoDB takes the same time to retrieve the same sample, it's feasible to calculate the average value of each sample to get Tab.3, and fit its trend line: $y = -34.07\ln(x) + 229.99$ and $R^2 = 0.8623$. There are some possible reasons that lead to a low reliability in R^2: (1) there are greater interference factors when sample size is small; (2) there are only 4 groups of samples. However, it's basically feasible to conclude that MongoDB's spatial retrieval efficiency remains unchanged in the same sample size, and with the increase of sample size, its performance also becomes increasingly stable.

TABLE III. ANALYSE OF MONGODB RESULT

Time(MongoDB)	Sample(10^4)	Time/Sample
253	1	253
1917	25	76.68
6094	100	60.94
23472	400	58.68

Further investigate the gap between ArcGIS and MongoDB in performance (Tab.4). With the increase of sample size, MongoDB's performance superiority becomes increasingly prominent. Especially when it's used to retrieve large points, it's increased from 5.69 times of 10 thousand-sample size to 10.66 times of 4 million-sample size; but when the same amount of points are retrieved, the sample size increases, and the performance of ArcGIS, however, is improved somewhat. Taking the retrieval of 9,801 points for example, in 10 thousand-sample size, the time consumed by it is 5.69 times of that by MongoDB, in 1 million-sample size, the time is 3.16 times, and in 4 million-sample size, the time is 2.76 times.

TABLE IV. ANALYSE OF MONGODB RESULT

points		ArcGIS(time)/MongoDB(time)
1w Sample	100	4.12
	361	4.67
	2500	5.52
	6400	6.10
	9801	5.69
25w Sample	100	3.26
	2500	2.59
	10000	3.35
	62001	4.96
	249001	9.01
100w Sample	9801	3.16
	39601	3.24
	250000	4.68
	640000	7.33
	998001	9.73
400w Sample	9801	2.76
	39601	2.83
	250000	3.18
	1000000	4.21
	3996001	10.66

V. CONCLUSIONS

This paper aims at the present mainstream geographic information system software, ArcGIS and MongoDB of NoSQL, and uses four sample point sets, which are 10 thousand, 250 thousand, 1 million, and 4 million. The specific test program is

correspondingly written for the development tool of the two software is used to compare their performance in retrieving the spatial point in a fixed area. 1) It's found that MongoDB's retrieval performance is much better than ArcGIS, and both retrieval performances are very regular. 2) MongoDB takes the same time to retrieve the same sample size, while ArcGIS takes stably increased time as retrieval quantity increases. 3) Their retrieval performance is stable, so both of them are mature system software.

The above results suggest that both products have a stable and reliable performance. As a mainstream NoSQL database product, MongoDB advances with the times and updates the version that supports spatial retrieval duly, greatly improving its retrieval performance, so that it has more wide application in the Internet and LBS. Although there are only limited kinds of spatial retrieval natively supported by, it can basically satisfy major geographic information services. By combining its intrinsic efficient retrieval performance, we believe that it can be used at a higher level.

What's selected for the test in this paper is the retrieval of simple point features, which is the intersection of the functions of both. ArcGIS is inferior to MongoDB in performance, but it still has a very stable performance and a powerful function. The tool set offered by ArcObjects can integrate many functions together, to provide a perfect service chain. The performance comparison between both products is also a microcosm of relational database and NoSQL. The relational database represented by ArcGIS is perfectly functional and powerful, while the NoSQL database represented by MongoDB shows a significant performance superiority [17].

REFERENCES

[1] K. Chodorow, "MongoDB: the definitive guide," O'Reilly Media, Inc. 2013.

[2] X.Y. Li and G.S. Luo, "Study on NoSQL theory and database," Telecommunications Science, 2012(12), pp. 23-30.

[3] Y.Q. Ma and Y.Y. Li, "NoSQL database technology in the mobile internet," Journal of Pingdingshan University, 2013(2), pp. 60-63.

[4] X.M. Lv and X.Y. Li, "Comparison analyze of NoSQL database and relational database," Microcomputer Applications, 2011(10): 55-58.70.

[5] F. Chang, et al., "Bigtable: A distributed storage system for structured data," ACM Transactions on Computer Systems (TOCS), 26(2), 2008, pp. 4.

[6] S. Tian, "A design and implementation of large-scale remote sensing data storage system base on MongoDB and HDFS," ZheJiang Unversity, 2013, pp.94.

[7] Z. Quan, J. Song, "Research of database base on NoSQL," Radio Engineering, 2013(09),pp.4-6/27.

[8] B. Sadoun, and O. Al-Bayari, "LBS and GIS technology combination and applications," Computer Systems and Applications. AICCSA'07. IEEE/ACS, 2007, pp. 578-583.

[9] M. Overmars, M.D. Berg, M.V. Kreveld and O. Schwarzkopf, "Computational Geometry: Algorithms and Applications," Berlin: Heidelberg, 1997.

[10] R. D, et al., "Survey on NoSQL for management of big data," Journal of Software, 2013(08), pp. 1786-1803.

[11] F. Wu, "Key technology research of LBS data management system based on MongoDB," Bulletin of Surveying and Mapping, 2014(07), pp.121-124.

[12] G. Wang, "MongoDB database of applied research and program optimization MongoDB," China Science and Technology Information, 2011(20), pp.93-94/96.

[13] Y. Zhang, Q. Gan and J. Liu, "Research on storage model of geographic information shared data based on MongoDB," Surveying and Mapping, 2014(04), pp. 147-150/172.

[14] X. Guo and H. Zhang, "The third geographica ldata model-geodatabase and its implement," Journal of Taiyuan University of Science and Technology, 2008(01), pp.5-8.

[15] Y. Liu and Y. Jiang, "The study of geodatabase based on object-oriented technology and standard relational database," Geomatics & Spatial Information Technology, 2009(04), pp.139-142.

[16] Y. Zhou, et al., "Design & exploitation of management system of spatial data based on ArcObject," Geospatial Information, 2006(02), pp.50-51.

[17] R. Cattell, "Scalable SQL and NoSQL data stores," ACM SIGMOD Record, 39(4), 2011, pp. 12-27.