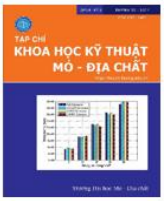
Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất
Trang điện tử: http://tapchi.humg.edu.vn



THÔNG TIN KHOA HỌC

Tổng quan về dữ liệu không gian địa lý lớn

Trần Mai Hương [1,*], Nguyễn Trường Xuân [1], Nguyễn Thị Mai Dung [1]

[1] *Khoa Công nghệ Thông tin, Trường Đại học Mỏ Địa chất, Việt Nam*

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 18/07/2017
Chấp nhận 20/8/2017
Đăng online 30/10/2017

Từ khóa:
Dữ liệu lớn
Dữ liệu không gian địa lý lớn
MapReduce
Xử lý dữ liệu không gian địa lý lớn

TÓM TẮT

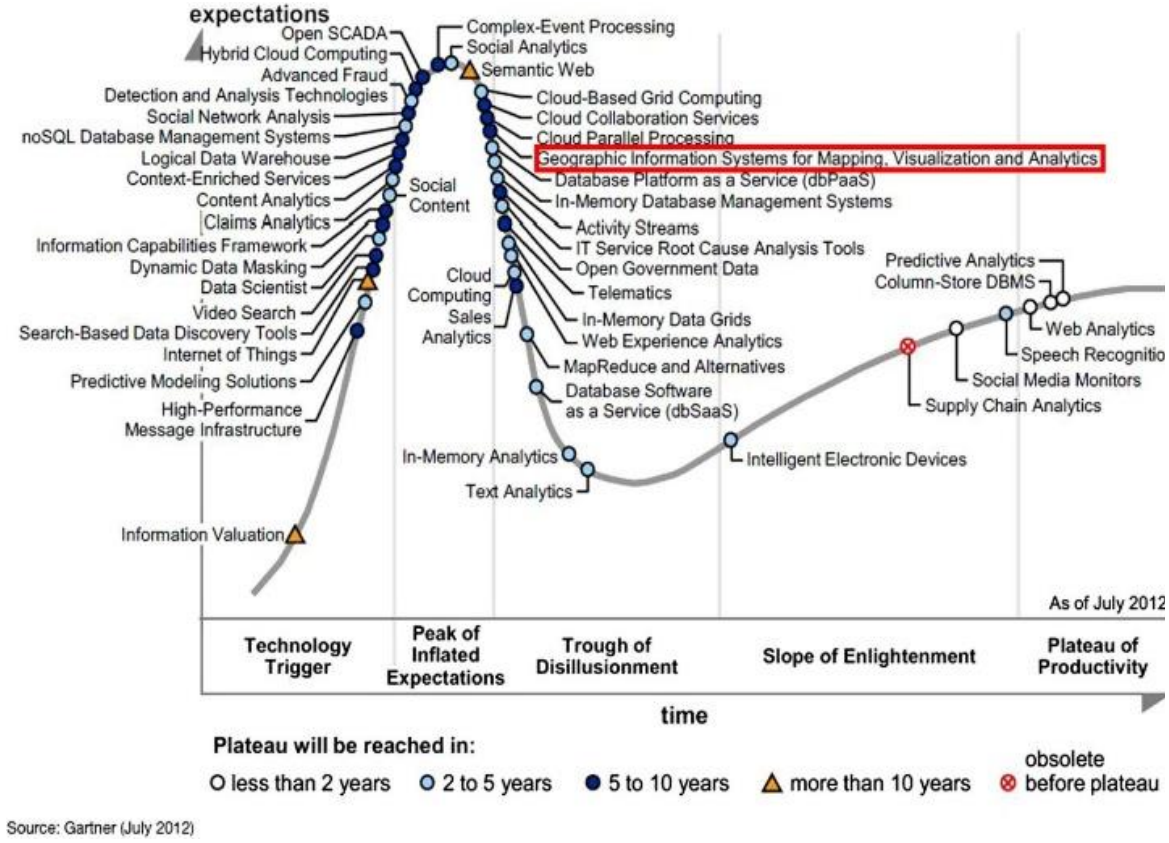
Dữ liệu không gian địa lý lớn(geospatial big data) được coi là các bộ dữ liệu không gian vượt quá khả năng xử lý của hệ thống máy tính hiện tại. Một phần quan trọng của dữ liệu lớn là dữ liệu được gắn với vị trí địa lý và dung lượng cũng như sự đa dạng của dữ liệu đó đang gia tăng nhanh chóng, ít nhất bằng 20% mỗi năm. Bài báo này cung cấp tổng quan về dữ liệu không gian địa lý lớn bao gồm: định nghĩa, các tính năng chính và các thành phần cần được hỗ trợ trong một hệ thống để xử lý dữ liệu này một cách hiệu quả. Ngoài ra bài báo cũng trình bày một số thách thức trong quá trình thu thập, xử lý dữ liệu không gian địa lý lớn.

1. Mở đầu

Chúng ta đang sống trong kỷ nguyên của 'Dữ liệu lớn- Big data'. Sự gia tăng nhanh chóng của dữ liệu lớn, đến từ nhiều nguồn khác nhau như các hệ thống cảm biến, hệ thống nhắn tin và mạng xã hội hệ thống vệ tinh, hệ thống định vị toàn cầu… Theo số liệu thống kê, kính viễn vọng không gian tạo ra dữ liệu không gian hàng tuần lên đến 150GB, các thiết bị y tế tạo ra các hình ảnh không gian (tia X) với tốc độ 50PB mỗi năm, một kho lưu trữ các hình ảnh vệ tinh của NASA có trên 500 TB và tăng mỗi ngày 25GB, trong khi có 10 triệu bài viết được gắn thẻ địa lý từ Twitter mỗi ngày chiếm khoảng 2% toàn bộ Twitter firehose. Theo ước tính của Hiệp hội quản lý thông tin địa không gian toàn cầu United Nations Initiative on Global Geospatial Information Management (UN-GGIM) có khoảng 2,5 tỷ Gigabyte dữ liệu được tạo ra mỗi ngày mà phần lớn trong đó đều liên quan tới vị trí địa lý. Các nhà khoa học thường gọi loại dữ liệu này là dữ liệu không gian địa lý. Theo nghiên cứu tổng hợp các loại dữ liệu của Gartner - công ty nghiên cứu và tư vấn về công nghệ thông tin, dữ liệu không gian địa lý phát triển mạnh vào năm 2012 (Lapkin, 2012) (Hình 1). Dữ liệu lớn, bao gồm cả dữ liệu không gian địa lý lớn đem đến rất nhiều lợi ích cho sự phát triển xã hội như là cung cấp thông tin phục vụ việc theo dõi sự thay đổi khí hậu, theo dõi dịch bệnh, ứng phó với thiên tai, giám sát các cơ sở hạ tầng chính, giao thông vận tải…

Sự gia tăng bùng nổ của dữ liệu không gian địa lý khiến cho dung lượng của dữ liệu đã vượt quá khả năng lưu trữ xử lý của các hệ thống máy tính hiện tại (Xu and Yang, 2014), (Lee and Kang, 2015). Vấn đề đặt ra là cần tìm một phương pháp xử lý khối dữ liệu địa lý lớn sao cho phù hợp. Vì thế trong bài báo này chúng tôi xem xét một số nghiên cứu hiện tại trong lĩnh vực dữ liệu không gian

*Tác giả liên hệ
E-mail: tranmaihuong@humg.edu.vn



Hình 1. Chu kỳ biến động của các loại dữ liệu (Nguồn: Gartner).

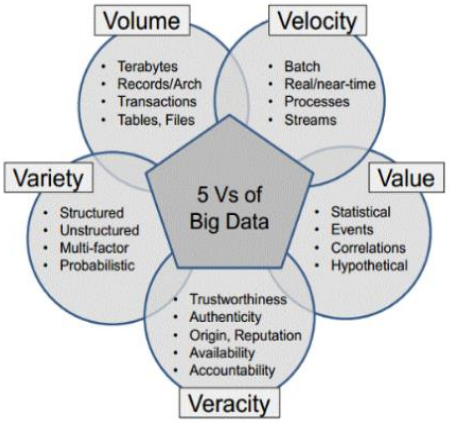
địa lý lớn để cung cấp một cái nhìn tổng quan về loại dữ liệu đặc biệt này.

Chúng tôi trình bày trong mục 2 định nghĩa và tính năng chính của dữ liệu không gian địa lý lớn. Công nghệ để xử lý dữ liệu không gian địa lý được nêu ở phần 3. Sau đó, chúng tôi tóm tắt một số thách thức trong việc xử lý dữ liệu không gian địa lý lớn trong phần 4. Mô hình xử lý dữ liệu không gian địa lý lớn được trình bày trong phần 5. Cuối cùng, phần 6 là kết luận bài báo.

2. Định nghĩa và các đặc trưng cơ bản của dữ liệu không gian địa lý lớn

2.1. Định nghĩa

Trên thế giới có nhiều định nghĩa về dữ liệu lớn. Vào năm 2001, nhà phân tích Doug Laney của hãng Gartner- được xem như là người đầu tiên đưa ra khái niệm về dữ liệu lớn. Gartner đã đưa ra định nghĩa dữ liệu lớn. - đó là dữ liệu có dung lượng lớn, tốc độ nhanh và có nhiều định dạng khác nhau mà các phương pháp truyền



Hình 2. Mô hình 5Vs.

thốngkhông đủ các ứng dụng để xử lý dữ liệu này. Gartner cho rằng dữ liệu lớn có thể được mô tả bằng mô hình "5Vs" (Hình 2) : Volume (Dung lượng lưu trữ), Velocity (Tốc độ xử lý), Variety (Đa dạng chủng loại), Veracity (Độ chính xác), Value (Giá trị thông tin). (Beyer and Laney, 2012).

Dữ liệu không gian địa lý lớn vẫn luôn được coi là dữ liệu lớn. Bài báo của Hiệp hội kỹ thuật quốc tế về Đo ảnh và Viễn thám (ISPRS) "Lý thuyết và các phương pháp xử lý dữ liệu địa lý lớn: đánh giá và định hướng nghiên cứu" (Li and et al., 2016) đã đưa ra những phân tích: Dữ liệu không gian địa lý mô tả những sự vật, hiện tượng có liên quan đến không gian địa lý, thường có vị trí tọa độ trong một hệ thống tham chiếu không gian. Dữ liệu này được thu nhận từ hệ thống vệ tinh toàn cầu (GNSS), mạng cảm biến, công nghệ radar và Lidar, công cụ Google Earth, Google Map…

2.2. Các đặc trưng của dữ liệu không gian địa lý lớn

Dữ liệu không gian địa lý thể hiện được ít nhất một trong số ba tiêu chuẩn 3Vs của dữ liệu lớn và được định nghĩa phù hợp với loại dữ liệu đặc biệt này (Li and et al., 2016).

(1) Khối lượng (Volume): Có khoảng 500 vệ tinh trên toàn cầu và một số vệ tinh đã hoạt động trong nhiều thập kỷ qua (ví dụ, các vệ tinh Landsat đã phục vụ gần 30 năm), thì dữ liệu hình ảnh vệ tinh nhận được đã trở thành số lượng rất lớn. Độ phân giải của ảnh càng cao thì dung lượng của chúng càng lớn(Dasgupta, 2013). Kích thước của dữ liệu không ngừng tăng lên tới terabyte và petabytes khiến việc lưu trữ và phân tích bằng các công cụ CSDL truyền thống rất khó khăn. Với công nghệ dữ liệu địa không gian lớn chúng ta có thể lưu trữ và sử dụng những tập dữ liệu này với các hệ thống phân tán.

(2) Tốc độ (Velocity): Dữ liệu ảnh có độ phân giải cao được thu thập liên tục thông qua các vệ tinh giám sát thời gian thực. Dữ liệu từ các nguồn khác (ví dụ như từ các cảm biến từ hệ thống định vị dẫn đường GNSS…) đòi hỏi quá trình xử lý dữ liệu phải đạt được tốc độ cao tương ứng.

(3) Sự đa dạng(Variety): dữ liệu địa không gian đến từ nhiều nguồn khác nhau như dữ liệu bản đồ, dữ liệu ảnh, dữ liệu text có gắn vị trí địa lý, có cấu trúc và không có cấu trúc, dữ liệu raster và vecter tất cả các loại dữ liệu khác nhau với những cấu trúc phức tạp đòi hỏi mô hình, cấu trúc và hệ quản trị cơ sở dữ liệu mới để quản lý dữ liệu hiệu quả hơn. Ví dụ: sử dụng NoSQL.

Ngoài ra, còn có các đặc tính 'Vs' khác được đề xuất để định nghĩa dữ liệu không gian địa lý lớn, chẳng hạn như giá trị (value), tính xác thực (veracity) và mô hình hóa (visualization) (Li S.Dragicevic S. và nhóm tác giả, 2016), (Fromm H và Bloehdorn S, 2014)

(4) Value (giá trị): "giá trị" cũng là một đặc điểm quan trọng dữ liệu địa không gian lớn. Việc tiếp cận được dữ liệu lớn sẽ chẳng có ý nghĩa gì nếu chúng ta không chuyển được chúng thành những thứ có giá trị. Khoa học công nghệ đã có các bước tiến lớn để quản lý và xử lý phần lớn các thông tin thiết yếu (phần có giá trị) từ tập đa dạng dữ liệu không gian địa lý (Tao and et al., 2013). Tuy nhiên, sẽ rất khó để xác định giá trị của các tập dữ liệu dữ liệu lớn và phức tạp. Đối với dữ liệu không gian địa lý với dung lượng lớn thì không thể và không cần thiết để trích xuất tất cả các thông tin mà chỉ cần tìm ra những dữ liệu có giá trị theo các ứng dụng về xử lý dữ liệu cụ thể.

(5) Veracity (Chính xác): dữ liệu địa không gian với rất nhiều dạng thức khác nhau vì vậy chất lượng và tính chính xác của dữ liệu rất khó kiểm soát. Công nghệ dữ liệu lớn và phân tích dữ liệu ngày nay cho phép chúng ta làm việc với những loại dữ liệu này. Tuy nhiên, khối lượng lớn thường đi kèm với việc thiếu chính xác và chất lượng của dữ liệu.

(6) Mô hình hóa (Visualization): cung cấp các bước quan trọng để có thể đem tư duy của con người vào để phân tích dữ liệu. Quá trình biểu diễn dữ liệu sẽ giúp các nhà phân tích tìm ra những kiểu mẫu, từ đó tìm ra được những giả thuyết mới và những cách thức hiệu quả để giới hạn dữ liệu và phân tích, tính toán chúng dễ dàng hơn. Mô hình hóa các dữ liệu cũng giúp những người sử dụng sau cùng dễ dàng nắm bắt được những loại dữ liệu và mối quan hệ của chúng một cách nhanh chóng thông qua những cách thức xây dựng mô hình đó.

3. Công nghệ xử lý dữ liệu không gian địa lý lớn

Dữ liệu không gian địa lý lớn sinh ra đòi hỏi phải có phương pháp mới để lưu trữ và phân tích nó. Một số công nghệ thường được sử dụng để xử lý dữ liệu lớn như Hadoop, MapReduce và phương pháp xử lý song song trên các hệ dữ liệu phân tán.

(1) Máy tính song song (Parallel Computing) - là tập hợp các bộ xử lý kết nối với nhau theo một kiến trúc xác định để cùng hợp tác hoạt động và trao đổi dữ liệu. Nó bao gồm việc xử lý dữ liệu trên nhiều máy cùng một lúc, mỗi máy chạy hệ điều hành, bộ nhớ, tốc độ tính toán và hoạt động trên các phần khác nhau của dữ liệu Do đó tính toán

song song giúp giảm thời gian để phân tích các dữ liệu lớn. (Bação, 2006), (Aji and et al., 2013)

(2) Hệ thống tệp tin phân tán (Distributed File System- DFS) - là một giải pháp cho phép người quản trị tập trung các dữ liệu nằm rời rạc trên các file server về một thư mục chung và thực hiện các tính năng replicate nhằm đảm bảo dữ liệu luôn sẵn sang khi có sự cố về file server

(3) Apache Hadoop - là một framework nguồn mở viết bằng Java cho phép phát triển các ứng dụng phân tán để xử lý bộ dữ liệu lớn một cách miễn phí.Một nền tảng ứng dụng hỗ trợ các ứng dụng phân tán với dữ liệu rất lớn.

- Hadoop cho phép các ứng dụng có thể làm việc với hàng ngàn node khác nhau với hàng petabyte dữ liệu. Hadoop lấy được phát triển dựa trên ý tưởng từ các công bố của Google về mô hình MapReduce và hệ thống file phân tán Google File System (GFS).

(4) Data Intensive Computing (Data Intensive Computing) - Là một hệ thống máy tính sử dụng phương pháp tính toán song song dữ liệu để xử lý dữ liệu lớn. Hệ thống này này dựa trên nguyên tắc sắp xếp dữ liệu và các chương trình hoặc thuật toán được sử dụng để thực hiện tính toán. Hệ thống song song và phân tán các máy tính độc lập kết nối với nhau làm việc như là một nguồn tài nguyên máy tính tích hợp duy nhất được sử dụng để xử lý / phân tích dữ liệu lớn.

4. Mô hình xử lý dữ liệu lớn

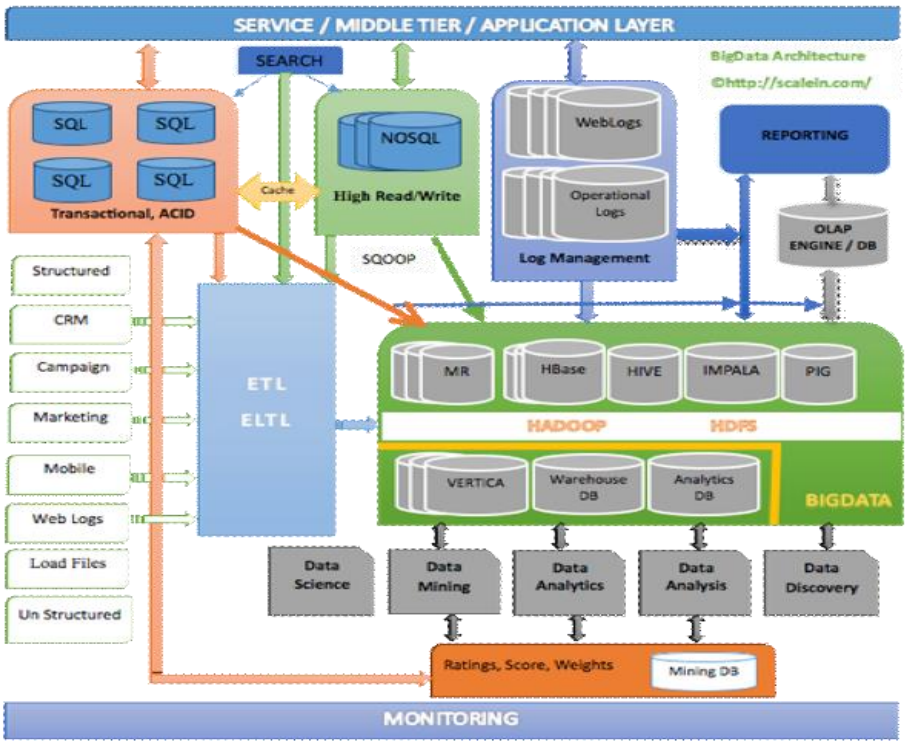
Dữ liệu không gian địa lý lớn được xử lý thông qua 4 giai đoạn(Hình 3): thu thập (acquire), tổ chức (organize), phân tích (analyze), quyết định (decide). (Horey and et al., 2012), (Ghemawatand et al., 2003), (Dean and et al., 2008).

- Giai đoạn thu thập: hầu hết đã có giải pháp, ví dụ: Oracle đưa ra NoSQL Database, Google có Google BigTable…

- Giai đoạn tổ chức: có thể lưu trữ dữ liệu ở dạng phân tán, song song… nhưng phổ biến nhất vẫn là Hadoop/MapReduce.

- Giai đoạn phân tích: với các dữ liệu truyền thống, các công ty lớn đều đã có giải pháp. Ví dụ: racle có Oracle Data warehousing, IBM có InfoSphere warehouse…

- Giai đoạn quyết định: dựa vào các thông tin được phân tích sẽ đưa ra các quyết định giải pháp kinh doanh kịp thời



Hình 3. Mô hình xử lý dữ liệu lớn.

5. Một số thách thức cho dữ liệu không gian địa lý lớn

Việc phân tích và xử lý dữ liệu không gian địa lý lớn đang đối mặt với nhiều thách thức - Quy mô dữ liệu: Thách thức về dung lượng và quy mô lớn chưa từng có của dữ liệu lớn đòi hỏi các công cụ quản lý và khai phá dữ liệu phải được cải tiến, nâng cấp cho phù hợp. Chúng ta có thể áp dụng 2 hướng tiếp cận sau trong khai phá dữ liệu lớn: (1) Điện toán đám mây kết hợp với tính toán song song; (2) tương tác người dùng: dựa trên giao diện đồ hoạ người dùng - GUI (Graphic User Interface) hoặc dựa trên ngôn ngữ tự nhiên, hướng tiếp cận này giúp việc tương tác giữa người dùng và hệ thống trở nên nhanh chóng và hiệu quả.

- Tốc độ xử lý trong các yêu cầu thời gian thực: Tốc độ/tính chuyển động liên tục của dữ liệu thực sự là một thách thức lớn. . Các kĩ thuật khai phá dữ liệu lớn phải có khả năng truy cập dữ liệu nhanh các dòng dữ liệu (data stream) - một định dạng phổ biến của dữ liệu lớn. Hệ thống xử lý dữ liệu phải hoàn thành việc xử lý/khai phá dòng dữ liệu đó trong một thời gian nhất định, bởi nếu không thì kết quả xử lý/khai phá đó trở nên ít có giá trị hoặc thậm chí là không có giá trị. Chẳng hạn, ứng dụng đòi hỏi chạy theo thời gian thực như dự đoán động đất … Tốc độ khai phá dữ liệu phụ thuộc vào hai yếu tố chính: thời gian truy cập dữ liệu (được xác định chủ yếu bởi hệ thống lưu trữ dữ liệu) và hiệu quả của các thuật toán khai phá dữ liệu. Chìa khóa giải quyết vấn đề này là việc sử dụng các chương trình lập chỉ mục tiên tiến. Trong đó, cấu trúc chỉ số đa chiều (Multidimensional index structure) đặc biệt hữu ích cho dữ liệu lớn. Một số nghiên cứu về chỉ mục đã công bố như: kết hợp RTree và KD-tree và gần đây là FastBit (phát triển bởi nhóm nghiên cứu ở LBNL). Tuy nhiên, hiện nay lập chỉ mục cho dữ liệu lớn vẫn là một trong những thách thức lớn nhất đối với cộng đồng nghiên cứu.

- Nền tảng xử lý dữ liệu không gian địa lý lớn: mặc dù Hadoop đã trở thành một trụ cột trong nền tảng phân dữ liệu lớn nhưng nó vẫn còn trong giai đoạn phát triển, so với cơ sở dữ liệu quan hệ. Đầu tiên, Hadoop phải tích hợp với thời gian thực cho việc thu thập và truyền dữ liệu lớn, và cung cấp xử lý nhanh hơn dựa trên các mô hình xử lý hàng loạt. Thứ hai, Hadoop nên cung cấp một giao diện lập trình ngắn gọn, và ẩn những tiến trình xử lý phức tạp bên dưới. Thứ ba, trong những hệ thống Hadoop lớn, số lượng máy chủ lên hàng ngàn, thậm chí hàng trăm ngàn, nghĩa là năng lượng tiêu thụ đáng kể. Vì vậy, Hadoop nên có cơ chế sử dụng năng lượng hiệu quả.

- Bảo mật dữ liệu và quyền riêng tư: là vấn đề rất quan trọng. Một số ví dụ trong thực tế cho thấy, không chỉ thông tin cá nhân người tiêu dùng, thông tin mật của các tổ chức mà ngay cả các bí mật an ninh quốc gia cũng có thể bị xâm phạm. Do vậy, giải quyết các vấn đề an ninh dữ liệu bằng các công cụ kỹ thuật và các chính sách trở nên vô cùng cấp bách. Các nền tảng dữ liệu lớn nên cân bằng tốt giữa việc truy cập dữ liệu và xử lý dữ liệu.

Như vậy, dữ liệu không gian địa lý lớn ngày càng đóng vai trò quan trọng. Để giải quyết được bài toán xử lý dữ liệu không gian địa lý lớn, đòi hỏi cần tổng hợp nhiều công nghệ và kỹ thuật khác nhau. Mỗi công nghệ và kỹ thuật cần có thời gian nghiên cứu và phát triển để hoàn thiện. Do vậy, với dữ liệu không gian địa lý lớn, rất nhiều lợi ích nhưng cũng còn nhiều vấn đề và thách thức cần giải quyết.

- Tính chính xác và tin cậy: dữ liệu không gian địa lý lớn có thể đến từ nhiều nguồn khác nhau, có thể từ nguồn không tin cậy và không thể kiểm chứng. Vì vậy, kết quả khai phá dữ liệu lớn là một thách thức cần giải quyết. Xác thực dữ liệu và xác minh nguồn gốc dữ liệu là một phương pháp để giải quyết phần nào thách thức này. Đây cũng là một bước quan trọng trong toàn bộ quá trình khai phá tri thức. Do dữ liệu lớn có tính động (dynamic) cao nên hệ thống phân tích và quản lý dữ liệu lớn cũng phải cho phép các dữ liệu được quản lý trong đó thay đổi và phát triển. Khi dữ liệu có sự thay đổi, phát triển thì các độ đo độ tin cậy cần được thay đổi hoặc cập nhật. Do đó, các độ đo này không nên được đặt cố định. Các nghiên cứu đã chỉ ra rằng, phương pháp học bán giám sát với dữ liệu thực tế (semi-supervised - học với tập dữ liệu huấn luyện gồm cả dữ liệu đã được gán nhãn và dữ liệu chưa được gán nhãn) có thể cung cấp độ chính xác và độ tin cậy cao hơn đối với các nguồn dữ liệu khác. Do dữ liệu lớn có tính động lớn nên "nguồn gốc dữ liệu" (data provenance) là thành phần không thể thiếu của bất kỳ hệ thống xử lý dữ liệu lớn nào. Nguồn gốc dữ liệu góp phần trực tiếp vào độ chính xác và tin cậy của kết quả khai phá dữ liệu. Tuy nhiên, thông tin về nguồn gốc dữ liệu không phải lúc nào cũng có sẵn hoặc được ghi

chép đầy đủ. Đây cũng là một thách thức của khai phá dữ liệu lớn. (Elwood S và nhóm tác giả, 2013), (Tang J C và nhóm tác giả, 2011), (Flanagin A J và nhóm tác giả, 2011)

- Sự tương tác: Sự tương tác là một vấn đề quan trọng trong khai phá dữ liệu lớn. Nó là tính năng của một hệ thống khai phá dữ liệu cho phép người dùng tương tác một cách nhanh chóng và đầy đủ (bao gồm phản hồi/can thiệp/hướng dẫn từ người dùng). Sử dụng thông tin phản hồi/hướng dẫn từ người dùng có thể giúp thu hẹp khối lượng dữ liệu, đẩy nhanh tốc độ xử lý, tăng khả năng mở rộng của hệ thống. Đồng thời, hệ thống tương tác cho phép người dùng có khả năng hình dung, đánh giá hoặc đánh giá trước cũng như giải thích kết quả khai phá trung gian và cuối cùng.

6. Kết luận

Dữ liệu không gian địa lý là một phần quan trọng của dữ liệu lớn, đã và đang được ứng dụng rộng rãi trong nhiều lĩnh vực. Tùy thuộc vào các nguồn và phương pháp thu thập khác nhau, dữ liệu không gian địa lý có thể được xác định theo các phạm vi khác nhau. Bên cạnh các đặc tính '3Vs' (và các 'Vs' khác) của dữ liệu lớn, thì dữ liệu không gian địa lý lớn có những tính năng độc đáo của riêng nó.

Việc nghiên cứu và ứng dụng công nghệ vào xử lý dữ liệu không gian địa lý lớn cần được đầu tư và quan tâm hơn nữa để nâng cao năng lực đội ngũ nghiên cứu và ứng dụng trong các bài toán xử lý dữ liệu địa lý lớn cũng như xử lý các loại dữ liệu không gian phi cấu trúc trong tương lai.

Tài liệu tham khảo

Dasgupta, A., 2013. *Big data: the future is in analytics.* http://www.geospatialworld.net/Magazine/MArticleView.aspx?aid=30512, Apr. 2013, Geospatial World.

Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., & Saltz, J., 2013. *Hadoop GIS: a high performance spatial data warehousing system over mapreduce*. Proceedings of the VLDB Endowment, 6(11), 1009-1020.

Bação, F. L., 2006. *Geospatial Data Mining. ISEGI*, New University of Lisbon.

Beyer, M. A., Laney, D., 2012. *The Importance of 'big data': a Definition*. Gartner.

Bakillah, M., Lauer, J., Liang, S. H. L., Zipf, A., Arsanjani, J. J., Mobasheri, A. and Loos, L., 2014. Exploiting big VGI to improve routing and navigation services: *Big Data Techniques and Technologies in Geoinformatics* ed. H A Karimi (CRC Press).

Bhosale, H. S., Gadekar, D. P., 2014. *A Review Paper on Big Data and Hadoop*.

Borthakur, D., 2007. *The hadoop distributed file system: Architecture and design*. Hadoop Project Website, 11.

Dunren Che, Mejdl Safran, and Zhiyong Peng, 2013. *From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, Database Systems for Advanced Applications,* pp 1-15, Springer Berlin Heidelberg.

Elwood, S., Goodchild, M. F., Sui, D., 2013. *Prospects for VGI Research and the Emerging Fourth Paradigm Crowdsourcing Geographic Knowledge: Volunteered Geographic* Information (VGI) in Theory and Practice ed D Sui, S Elwood and M F Goodchild (New York: Springer). 361-375.

Fromm H and Bloehdorn S 2014 *Big Data-Technologies and Potential Enterprise-Integration* ed G Schuh and V Stich (Berlin: Springer) pp 107-124.

Horey, J., Begoli, E., Gunasekaran, R., Lim, S., and Nutaro, J., 2012. *Big data platforms as a service: Challenges and approach*. Proceedings of the 4th USENIX conference on Hot Topics in Cloud Ccomputing, pp. 16-16, USENIX Association.

Laney, D. 2001. *3D Data Management: Controlling Data Volume, Velocity and Variety.* (Gartner) Hyderabad, March 28-April 1

Lee, J. G., Kang, M., 2015. *Geospatial big data: challenges and opportunities.* Big Data Res.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., 2016. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*. 115. 119-133.

Tao, X., Hu, X., and Liu, Y., 2013. *Review on Big Data Research* J. System Simulation 25 (Supplement) 142-146.

Tang, J. C., Cebrian, M., Giacobe, N. A., Kim, H. W., Kim, T. and Wickert, D. B., 2011. *Reflecting on the Darpa Red Balloon Challenge Communications of the ACM* 54 4 pp 78-85

Xu, C., and Yang, C., 2014. *Introduction to big geospatial data research* Ann. GIS 20 4 pp 227-232.

ABSTRACT

Geospatial big data: an overview

Huong Mai Tran [1], Xuan Truong Nguyen [1], Dung Mai Thi Nguyen [1]

[1] *Faculty of Information Technology, Hanoi University of Mining and Geology, Vietnam*

Geospatial data is considered the largest data sets space exceeds the capabilities of the current computer system. An important part of the large data is data that is associated with the geographical location and size as well as the diversity of data that are rising quickly, at least by 20% each year. This paper provides an overview on geospatial big data include: definitions, the main features and components need to be supported in a system for handling this data efficiently. In addition, the challenges and opportunities of geospatial big data processing also outlined in this paper.

Keywords: Big data, Geospatial big data, MapReduce, Geospatial big data processing.