DỮ LIỆU LỚN VÀ DỮ LIỆU LIÊN KẾT MỞ: MỘT SỐ HƯỚNG TIẾP CẬN

Phạm Xuân Hậu

Tóm tắt. *Dữ liệu lớn (Big Data) và dữ liệu liên kết mở (Linked Open Data) là các vấn đề nghiên cứu mới và thú vị đang được các nhà nghiên cứu, các doanh nghiệp và các nhà phát triển ứng dụng quan tâm. Trong phạm vi bài báo này, tôi xin trình bày những vấn đề liên quan và các hướng tiếp cận đang được tập trung nghiên cứu trong lĩnh vực này.*

Từ khóa: Dữ liệu lớn, dữ liệu liên kết mở, mạng xã hội, dữ liệu, liên kết.

1. GIỚI THIỆU

Với sự bùng nổ và phát triển không ngừng của Internet, các kho thông tin và dữ liệu đang phát triển theo cấp số nhân và các kiểu dữ liệu ngày càng đa dạng phong phú. Ngày nay, với sự phát triển của công nghệ web, các kho dữ liệu không chỉ dừng ở terabyte, petabyte mà nó đã lên đến exabyte, zettabyte. Các dữ liệu có cấu trúc, không có cấu trúc, các kho dữ liệu phân tán hay tập trung, các dữ liệu không rõ ràng hay các dữ liệu ẩn và với kích thước rất lớn được gọi là dữ liệu lớn (Big Data-BD). Bên cạnh đó, với sự phát triển của các mạng xã hội, các hệ thống mở đã cho phép các dữ liệu được kết nối với nhau nhằm chia sẻ và cộng tác từ nhiều nguồn khác nhau được gọi là các dữ liệu mở (Linked Open Data-LOD). Chúng đã mang đến những thách thức cho các nhà nghiên cứu và các nhà phát triển trong việc khai thác các dữ liệu và triển khai các dịch vụ trên các kho dữ liệu đó.

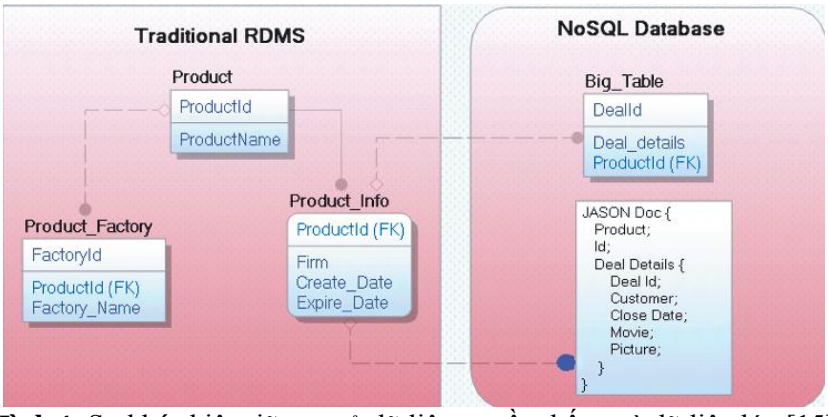
Trong phạm vi bài báo này, tôi sẽ trình bày tổng quan về BD, LOD, các thách thức và các cách tiếp cận mà các nhà nghiên cứu đang nghiên cứu.

1.1. Dữ liệu lớn (Big Data)

Dữ liệu lớn (BD) là một thuật ngữ mà khái niệm của nó chưa được trình bày rõ ràng và còn nhiều cách hiểu khác nhau về thuật ngữ này [1]. Tuy nhiên, nhìn chung dữ liệu lớn được hiểu là các dữ liệu có kích thước cực lớn, không có cấu trúc, các mối quan hệ không rõ ràng, tổ chức phức tạp, chúng tồn tại trên mạng xã hội, các thiết bị thông minh, các cảm biến, các dữ liệu thời gian thực,… Các mô hình dữ liệu truyền thống thường tập trung vào việc phân tích, xử lý các dữ liệu có cấu trúc, rõ ràng và kích thước nhỏ. Tuy nhiên để xử lý các BD thì các mô hình truyền thống gặp nhiều khó khăn [15]. Với xu thế hiện nay, các công ty, các tổ chức đang bắt đầu tiếp cận để hiểu và khai thác các lợi ích mà các thông tin từ BD mang lại để từ đó có những hoạch định chiến lược kinh doanh phù hợp. Với những kết quả mang lại bất ngờ đã làm thay đổi quan niệm tiếp cận thông tin của rất nhiều đối tượng để mang lại nhiều lợi ích trong kinh doanh [2,3].

Theo thống kế, đến năm 2003 đã có 5 exabytes (10^{18} bytes) dữ liệu được tạo ra bởi các tổ chức và cá nhân người dùng. Và hiện nay chỉ cần 2 ngày là con người đã tạo ra được khối lượng dữ liệu tương đương. Dữ liệu toàn thế giới ước khoảng 2.72 zettabytes (10^{21} bytes) dữ liệu vào

năm 2012, cứ mỗi 2 năm thì kích thước của nó tăng gấp đôi và năm 2015 ước đạt khoảng 8 zettabytes. Theo dự báo của IBM thì mỗi ngày có khoảng 2.5 exabytes dữ liệu được tạo ra (hằng ngày có khoảng 6 tỷ cuộc gọi, 10 tỷ tin nhắn được gửi đi). Tính trung bình hiện nay mỗi máy tính cá nhân có thể chứa khoảng 500 gigabytes (10^9 bytes) dữ liệu và điều này có nghĩa là chúng ta cần có khoảng 20 tỷ máy tính cá nhân để chứa tất cả dữ liệu trên toàn thể giới. Internet trở thành kênh chính để lưu chuyển dữ liệu [4,7].



Hình 1. Sự khác biệt giữa cơ sở dữ liệu truyền thống và dữ liệu lớn [15].

Có 3 đặc trưng quan trọng của BD đó là: tốc độ (velocity), dung lượng (volume) và đa dạng (variety). Tốc độ thể hiện yêu cầu về mặt thời gian cho quá trình xử lý từ các khối dữ liệu (batch) cho đến các luồng dữ liệu (stream) phải được đảm bảo. Dung lượng thể hiện kích cỡ của dữ liệu phải xử lý từ terabytes đến yottabytes. Đa dạng thể hiện định dạng dữ liệu phải xử lý rất đa dạng từ dạng có cấu trúc, bán cấu trúc và không có cấu trúc.

Google là một ví dụ điển hình cho mô hình BD (Google có một triệu server để lưu trữ, xử lý thông tin). Mục đích chủ yếu là hiểu được người sử dụng và cung cấp cho họ những thông tin mà họ tìm kiếm. Google đã sử dụng nhiều công cụ, thuật toán và kỹ thuật lưu trữ, xử lý dữ liệu (schema, RDF, microformats, PageRank [14],…). Thông tin trên mạng xã hội như Facebook, Twitter, Youtube,… cũng được xem xét như những BD. Bảng 1, cho chúng ta thấy một số thông tin về các mạng xã hội được thống kê bởi các nhà phát triển chúng.

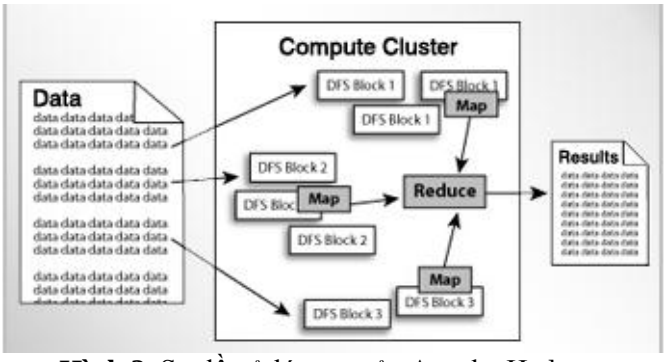
Bảng 1. Thống kê số liệu về mạng xã hội Facebook, Twitter, Youtube

Facebook	Twitter	Youtube
- 890 triệu người sử dụng hằng ngày (1.39 tỷ người hằng tháng) - 745 triệu người sử dụng trên di động hàng ngày (1.19 tỷ người trong tháng) - Hỗ trợ trên 77 ngôn ngữ - 2.7 tỷ like mỗi ngày	- 288 triệu người hàng tháng - 500 triệu Tweets được gửi trong 1 ngày - 80% sử dụng Twitter trên di động - Hỗ trợ 33 ngôn ngữ	- 1 tỷ người dùng - 300 giờ video/1 phút được tải lên - Có hằng trăm triệu giờ xem và hàng tỷ lượt xem mỗi ngày - Hỗ trợ 61 ngôn ngữ

- 300 triệu ảnh được tải mỗi ngày		

BD là các dữ liệu mà ở đó không thể dùng các công cụ, phương pháp thông thường để khai thác và phân tích chúng [1,2,3,4,7]. Vì vậy, một số công cụ (MapReduce, Hadoop) hỗ trợ cho việc xử lý các tập dữ liệu lớn phân tán và cơ sở dữ liệu NoSQL (Cassandra, MongoDB) cung cấp một cơ chế cho việc lưu trữ và thu hồi dữ liệu các tập dữ liệu lớn được nghiên cứu và phát triển [2,3].

- *Hadoop*[1]: là một phần mềm nền hỗ trợ các ứng dụng dữ liệu phân tán cho phép viết các ứng dụng xử lý nhanh các BD dưới dạng song song. Nó được thiết kế để có thể thích hợp cho việc tính toán và lưu trữ từ các máy chủ đơn cho đến hàng ngàn máy. Hadoop gồm có Hadoop kernel, Map/Reduce và HDFS (Hadoop Distributed File System).



Hình 2. Sơ đồ xử lý cụm của Apache Hadoop.

- *MapReduce*: được xem như là "trái tim" của Hadoop. Nó là một mô hình lập trình và thực hiện các xử lý cho các dữ liệu kích thước lớn. Chúng thực hiện dựa trên phương pháp chia để trị (divide and conquer). MapReduce chia ra 2 bước [17,7]: Map và Reduce (tương ứng trong kiến trúc của Hadoop là nút master và nút worker). Bước Map: nút master sẽ chia nhỏ ra các bài toán con và phân phối cho các nút worker. Bước Reduce: nút master sẽ tổng hợp các kết quả từ các bài toán con để đưa ra kết quả chung. Một sơ đồ xử lý của Apache Hadoop được mô tả ở Hình 2.

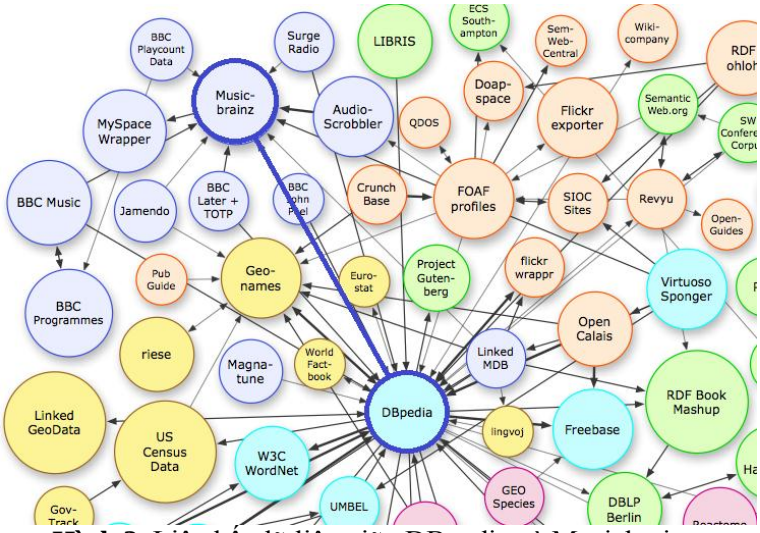
[1] http://hadoop.apache.org/

- *Cassandra*[2]: Cassandra là một cơ sở dữ liệu dùng cho việc lưu trữ phân tán và được phát triển cho việc khai thác dữ liệu trên mạng xã hội như Facebook, Twitter, Reddit. Cassandra được thiết kế để đảm bảo việc thực hiện nhanh, hiệu quả, tin cậy và có khả năng tái tạo và là một mô hình có tính mềm dẻo [2,16]. Một bảng trong cơ sở dữ liệu của Cassandra là một dạng phân tán nhiều chiều được định chỉ số bởi một khóa.
- *MongoDB*[3]: trong khi Cassandra là cơ sở dữ liệu hướng cột (column-oriented) thì MongoDB là một cơ sở dữ liệu hướng văn bản (document-oriented).

1.2. Dữ liệu liên kết mở (Linked Open Data)

Với sự phát triển nhanh chóng của công nghệ web, việc truyền tải, chia sẻ thông tin không còn mang nhiều yếu tố địa lý và ngôn ngữ. Các hệ thống mở được phát triển mà ở đó mỗi người dùng cũng là một nhà phát triển dữ liệu cho hệ thống [12,13].

Dữ liệu liên kết mở (LOD) được hiểu là các dữ liệu trên web được liên kết lại với nhau và chia sẻ với mọi người. Thuật ngữ LOD[4] được đề xuất thông qua dự án từ tháng 1/2007 và được hỗ trợ bởi *W3C Semantic Web Education and Outreach Group*. Tốc độ phát triển của LOD ngày càng cao thông qua các dự án của các tổ chức và các công ty với tiêu chí mọi người có thể tham gia để xuất bản các dữ liệu theo các nguyên tắc của LOD và liên kết với các dữ liệu đang tồn tại bởi các interlink [9].



Hình 3. Liên kết dữ liệu giữa DBpedia và Musicbrainz.

Hiện nay, việc tính toán kích thước chính xác của dữ liệu web là một thách thức vì thực tế việc tạo ra và xuất bản dữ liệu trên các hệ thống mở phải được thu thập trước khi tính toán [8,9]. Số lượng tập dữ liệu ngày càng nhiều, các nguồn liên kết càng mở rộng [22]. Tuy nhiên, chúng ta có thể thống kê được một vài con số thú vị về một số LOD như trong Bảng 2.

[2] http://cassandra.apache.org/
[3] http://www.mongodb.org/
[4] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Bảng 2. Thống kê số liệu về các LOD [25,26,27]

Wikipedia	DBpedia	LinkedMDB
4,732,709 bài đăng 35,232,726 trang 848,655 file 758,950,132 biên tập 24,243,323 người dùng 1,359 quản trị	4,580,000 đối tượng 1,445,000 người 735,000 địa danh 123,000 âm nhạc (album) 87,000 phim 19,000 game 241,000 tổ chức 125 ngôn ngữ	6,148,121 RDF 162,199 interlink 541,810 liên kết tham khảo (movie) 503,242 thực thể

Theo Bizer C. [11], liên kết dữ liệu trên web được hiểu là tạo các kết nối giữa các dữ liệu từ các nguồn khác nhau. Trong khi phần chính của siêu văn bản (hypertext) là các HTML (HyperText Markup Language), các văn bản kết nối với nhau bằng siêu liên kết (hyperlink) thì đối với liên kết dữ liệu dựa vào các văn bản có chứa dữ liệu trong định dạng RDF (Resource Description Framework). Liên kết dữ liệu được xây dựng dựa vào: URI (Uniform Resource Identifiers) và URL (Uniform Resource Locators). URL thể hiện mối quan hệ giữa các thực thể và toàn bộ văn bản được đinh vị trên trang web. URI thể hiện một kiểu để xác định bất kỳ thực thể nào tồn tại, chúng sử dụng cấu truc http:// thông qua giao thức HTTP (HyperText Transfer Protocol). Hình 3. mô tả sự kết nối giữa các LOD thông qua liên kết RDF và interlink.

Interlink là một khái niệm chỉ sự kết nối giữa giữa các nguồn dữ liệu khác nhau. *RDF* là một cơ chế để xác định sự tồn tại và nghĩa của một kết nối giữa các thực thể trong tập dữ liệu.

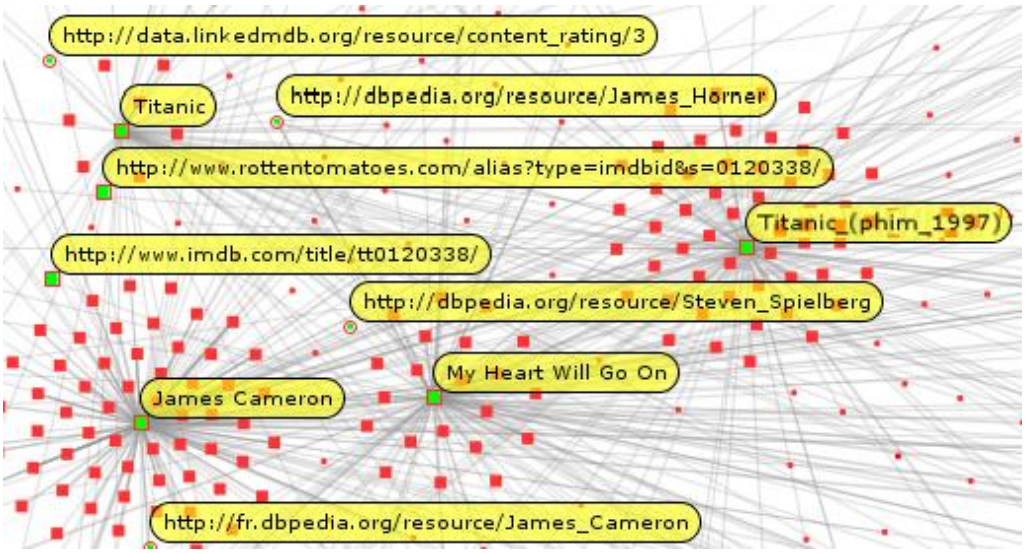
Mô hình RDF cung cấp một cấu trúc kiểu đồ thị mà ở đó cấu trúc và liên kết dữ liệu của tất cả mọi thứ đang tồn tại. RDF được định nghĩa là một bộ ba *<chủ thể, quan hệ, đối tượng>*. Chủ thể (subject) là một URI để xác định nguồn dữ liệu được mô tả. Đối tượng (object) có thể là một chuỗi, số hoặc một URI của một nguồn khác có liên quan đến chủ thể. Quan hệ (predicate) là một URI chỉ rõ mối quan hệ giữa chủ thể và đối tượng. Ta xét ví dụ sau:

Chủ thể: *http://dbpedia.org/page/Titanic_(1997_film)*

Quan hệ: *http://xmlns.com/foaf/0.1/homepage*

Đối tượng: *http://www.titanicmovie.com/*

RDF này là một mô tả của bộ phim "*Titanic*" có URI *http://dbpedia.org/page/Titanic_(1997_film)* trên DBpedia cho biết nhà sản xuất bộ phim có trang chủ là *http://www.titanicmovie.com/*. Hình 4 thể hiện việc liên kết giữa các interlink và các RDF giữa các LOD.



Hình 4. Interlink và RDF trong các LOD [13].

Để có thể truy xuất được các RDF và interlink trong LOD, một cơ chế, giao thức truy vấn được phát triển, đó là SPARQL. SPARQL là một giao thức và ngôn ngữ truy vấn RDF. SPARQL cung cấp các mô tả dữ liệu dưới dạng tri thức. Ngôn ngữ truy vấn SPARQL có thể được sử dụng để diễn tả các truy vấn đến các nguồn dữ liệu khác nhau. Bên cạnh đó, SPARQL cũng hỗ trợ các cú pháp và ngữ nghĩa cho truy vấn. Kết quả của truy vấn có thể được trình bày dưới định dạng RDF hoặc các định dạng khác [13]. Chúng ta xét ví dụ, để lấy thông tin về bộ phim "*Titanic*" trên DBpedia, chúng ta sẽ dùng truy vấn sau:

```
SELECT ?film ?film_label
WHERE
{
        ?film rdf:type dbpedia:Film .
        ?film rdfs:label ?film_label.
        FILTER (regex(?film, 'Titanic', 'I'))
}
```

Việc đa dạng hóa các kiểu dữ liệu và ngôn ngữ thể hiện chúng cũng gặp nhiều khó khăn trong việc kết nối chúng. Và các nhà phát triển đã đề xuất một số kiểu liên kết liên quan đến các thực thể ngôn ngữ như ILL hay IRI. *Inter-language link* (ILL) là liên kết ngôn ngữ đặc biệt từ một trang bất kỳ với một ngôn ngữ xác định đến một trang tương ứng được thể hiện ở một ngôn ngữ khác. Mỗi tập dữ liệu có thể được trình bày với các ngôn ngữ khác nhau. *Internationalized Resource Identifiers* (IRI) là các định danh để xác định các nguồn dữ liệu sử dụng các ký tự Unicode [10,11,13].

Hiện nay với sự phát triển không ngừng của LOD, có rất nhiều hệ thống được xuất bản. Một số hệ thống phổ biến hiện nay như hệ thống Wikipedia, DBpedia và LinkedMDB.

- Wikipedia[5]: là một hệ thống mở được xem như bách khoa toàn thư. Nó được phát triển bởi cộng đồng, bất kỳ ai cũng có thể trở thành một thành viên thông qua việc cập nhật, bổ sung, sửa chữa thông tin. Hệ thống chứa nhất nhiều chủ đề khác nhau với sự hỗ trợ của nhiều phiên bản ngôn ngữ.

- DBpedia[6]: là một hệ thống mở trên Internet. Nó chứa rất nhiều các dữ liệu (things) được mô tả dưới nhiều ngôn ngữ khác nhau được rút trích từ Wikipedia, được tổ chức lại một cách có hệ thống và có sự liên kết với các nguồn dữ liệu khác để mô tả thông tin rõ hơn cho các thực thể.

- LinkedMDB[7]: là một hệ thống mở về phim dựa vào các thông tin rút trích có được từ IMDB (Internet Movie Database). Nó chứa rất nhiều các liên kết đến các nguồn khác nhau có liên quan đến các thực thể (nội dung) của bộ phim.

2. CÁC THÁCH THỨC VÀ CÁC HƯỚNG TIẾP CẬN

BD và LOD là các lĩnh vực nghiên cứu mới nên nó đang đặt ra nhiều vấn đề thú vị cho các nhà nghiên cứu và phát triển. Trong mục này tôi trình bày một vài hướng tiếp cận mà tôi và các đồng sự đã nghiên cứu.

Như chúng ta đã biết, với kích thước cực lớn và cấu trúc phức tạp của BD đã đặt ra những thách thức cho người sử dụng cũng như các doanh nghiệp đó là: Làm thế nào để nhanh chóng tiếp cận, khai thác nhanh thông tin hữu ích trong một "núi khổng lồ" dữ liệu như thế? Làm sao để hiểu được chúng? Làm thế nào để đảm bảo chất lượng thông tin từ chúng? Cách trình bày ra sao? Việc phát triển ứng dụng, các chiến dịch quảng bá, truyền thông.

Một số hướng tiếp cận để khai thác thông tin từ BD dựa trên mạng xã hội đã được trình bày [14,19,20,22,23]. Trên mạng xã hội Twitter với số lượng tweet/retweet cực lớn, thông thường một tweet thường đi kèm với sự kiện [19,22]. Một cách tiếp cận để nhận biết các sự kiện theo dòng thời gian dựa trên luồng dữ liệu (Data Stream) đã được đề xuất [19]. Bên cạnh đó, việc khai thác, hiểu được đầy đủ ý nghĩa các mối quan hệ, thói quen của người sử dụng dựa trên mối quan hệ tương ứng giữa các tập dữ liệu phức hợp và các thuộc tính cụ thể được trực quan hóa được thể hiện trên hệ thống TweetScope [22].

Một số dữ liệu quan trong trên mạng xã hội thường được nghiên cứu và khai thác đó chính là thẻ (tag), bình luận (comment), thông tin người sử dụng và mối quan hệ giữa chúng, người sử dụng và các đối tượng khác. Trong [20,23], một nghiên cứu về việc truyền bá thông tin thông qua mạng xã hội đã được nghiên cứu. Một hệ thống thẻ mà ở đó người dùng có thể trao đổi các nguồn thông tin cũng như thông tin của họ (các thẻ) với người khác một cách dễ dàng. Giả sử

[5] http://wikipedia.org
[6] http://dbpedia.org
[7] http://linkedmdb.org

rằng, "nhịp đập" của mạng xã hội (social pulse) được hình thành dựa trên số lượng người sử dụng có gắn các thẻ đó. Vì vậy, chúng ta có thể khai thác được đầy đủ các mối quan hệ giữa các thẻ với nhau. Một nghiên cứu sử dụng các thể địa danh (geotag) trên Flickr[8] đã được trình bày trong [14]. Trong mỗi tập các ảnh được gắn các thẻ trong đó có thể có thẻ địa danh sẽ được phân tích và tính toán. Mục tiêu của chúng tôi là xếp hạng các địa danh trong tập dữ liệu đó. Bên cạnh đó, một ứng dụng hệ khuyến nghị đã được phát triển dựa trên các thông tin của người dùng và ngữ cảnh trên Facebook [24].

Với sự phát triển không ngừng của web và các kỹ thuật xử lý, tổ chức dữ liệu trên web các nguồn LOD ngày càng đa dạng và phong phú đang là thách thức cho các nhà nghiên cứu và phát triển, đó là: giao diện người dùng như thế nào? Tương tác, kết nối các mô hình ra làm sao? Kiến trúc các ứng dụng được thực hiện như thế nào? Các sơ đồ ánh xạ và hợp nhất dữ liệu, vấn đề bản quyền về dữ liệu (Licensing), sự tin tưởng (Trust), chất lượng (Quality) và các vấn đề liên quan của dữ liệu. Một vấn đề thách thức rất lớn trong LOD nữa đó là tính riêng tư (Privacy) của dữ liệu khi liên kết và chia sẻ trên hệ thống. Bên cạnh đó, cùng với các giá trị có nghĩa mà LOD mang lại cho người dùng, rất nhiều nghiên cứu và ứng dụng bằng cách khai thác các dữ liệu từ LOD. Các hướng tiếp cận của LOD tập trung vào: các trình duyệt trên nền tảng LOD, các máy tìm kiếm, các ứng dụng cụ thể [11].

Ngôn ngữ là một chủ đề rất thú vị trong LOD. Các web thông thường thì hệ thống chỉ có một ngôn ngữ xác định để trình bày, việc chuyển đổi giữa các ngôn ngữ được thực hiện qua các máy dịch, điều này gây hạn chế cho việc kết nối các dữ liệu. Tuy nhiên trong các hệ thống LOD thì việc dữ liệu được trình bày với nhiều ngôn ngữ lại được quan tâm, các dữ liệu được tổ chức dưới dạng các ontology hoặc dưới dạng các bộ từ điển nên việc truy xuất và trình bày dữ liệu không bị phụ thuộc vào văn hóa hay ngôn ngữ bản địa. Hệ thống đơn ngữ (Monolingual System): các dữ liệu được mô tả và trình bày chỉ bởi một ngôn ngữ. Vì vậy, việc trình bày các thông tin và tri thức liên quan đến dữ liệu gặp nhiều hạn chế. Đơn ngữ được mô tả như dạng một văn bản, các cuộc hội thoại. Hệ thống đa ngữ (Multilingual System): trong web 2.0, các dữ lệu có thể được hỗ trợ với nhiều ngôn ngữ trong nhiều hoàn cảnh khác nhau để đáp ứng nhu cầu của con người. Người dùng hoặc có thể hiểu một vấn đề được trình bày trong nhiều ngôn ngữ khác nhau hoặc có thể lấy dữ liệu mà không cần xét đến vấn đề ngôn ngữ. Các hệ thống đa ngôn ngữ có thể trình bày bất kỳ dữ liệu gì mà không phụ thuộc vào yếu tố ngôn ngữ, ví dụ như DBpedia hỗ trợ 125 ngôn ngữ. Khả năng đa ngôn ngữ của hệ thống có thể hỗ trợ người dùng một cách mềm dẻo và tiện lợi [13,21].

Hệ khuyến nghị là một trong những hướng nghiên cứu được nhiều người quan tâm và các công ty, các doanh nghiệp phát triển. Ở đó, các hệ thống cung cấp cho khách hàng các mặt hàng mà họ cần, chúng có thể hiểu được thói quen, sở thích và cả lịch sử mua bán của khách hàng.

[8] https://www.flickr.com

Một hướng tiệp cận hệ khuyến nghị dựa trên dữ liệu LOD đã được trình bày trong [12]. Trong đề xuất này, chúng tôi khai thác các dữ liệu từ nhiều nguồn khác nhau thông qua các interlink và các RDF để phát triển hệ thống không những khuyến nghị một kiểu mặt hàng (phim) mà còn mở rộng ra nhiều mặt hàng khác (sách, thời trang, âm nhạc).

3. KẾT LUẬN

Với xu hướng của công nghệ web hiện đại, cùng với vai trò và ý nghĩa hết sức to lớn của BD và LOD. Đây là 2 lĩnh vực nghiên cứu đang được các nhà khoa học và các doanh nghiệp quan tâm. Trong bài báo này, chúng tôi đã trình bày một cách tổng quan một số vấn đề liên quan đến chúng. Bên cạnh đó, chúng tôi cũng đã khảo sát, trình bày các thách thức mà chúng ta gặp phải, một số hướng tiếp cận mà chúng tôi đã và đang nghiên cứu.

TÀI LIỆU THAM KHẢO

[1] Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). *Analytics: The real-world use of big data*. IBM Institute for Business Value—executive report, IBM Institute for Business Value.

[2] Team, O. (2011). *Big Data Now: Current Perspectives*. O'Reilly Radar.

[3] Russom, P. (2011). *Big data analytics*. TDWI Best Practices Report, Fourth Quarter.

[4] Fan, W., & Bifet, A. (2013). *Mining big data: current status, and forecast to the future*. ACM SIGKDD Explorations Newsletter, 14(2), 1-5.

[5] Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

[6] Cavoukian, A., & Jonas, J. (2012). *Privacy by design in the age of big data*. Information and Privacy Commissioner of Ontario, Canada.

[7] Sagiroglu, S., & Sinanc, D. (2013, May). *Big data: A review*. In Collaboration Technologies and Systems (CTS), IEEE, pp. 42-47.

[8] Möller, K., Hausenblas, M., Cyganiak, R., & Handschuh, S. (2010). *Learning from linked open data usage: Patterns & metrics*.

[9] Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008, April*). Linked data on the web (LDOW2008)*. In Proceedings of the 17th international conference on World Wide Web, ACM, pp. 1265-1266.

[10] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007*). Dbpedia: A nucleus for a web of open data* . Springer Berlin Heidelberg, pp. 722-735.

[11] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far.

[12] Pham XH, Jung JJ and Takeda H. *Exploiting linked open data for attribute selection on recommendation systems.* In: the Proceedings of the 7th KES Conference on Agent and Multi-Agent Systems - Technologies and Applications (KES-AMSTA 2013), Vietnam, 2013, pp. 427–433.

[13] Pham XH, & Jung JJ. *Recommendation system based on multilingual entity matching on linked open data.* Journal of Intelligent and Fuzzy Systems, vol 27(2), 2014, pp.589-599

[14] Pham, X. H., Nguyen, T. T., Jung, J. J., & Hwang, D. (2014). Extending HITS algorithm for ranking locations by using geotagged resources. In Computational Collective Intelligence. Technologies and Applications. Springer International Publishing, pp. 332-341.

[15] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). *Data mining with big data.Knowledge and Data Engineering*, IEEE Transactions on, 26(1), pp. 97-107.

[16] Lakshman, A., & Malik, P. (2009, August). Cassandra: structured storage system on a p2p network. In Proceedings of the 28th ACM symposium on Principles of distributed computing, ACM.

[17] Chen, C. P., & Zhang, C. Y. (2014). *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data*. Information Sciences, 275, pp. 314-347.

[18] Bauer, F., & Kaltenböck, M. (2011). *Linked open data: The essentials.* Edition mono/monochrom, Vienna.

[19] Trung, D. N., Jung, J. J., Lee, N., & Kim, J. (2013). *Thematic analysis by discovering diffusion patterns in social media: an exploratory study with tweetScope*. In Intelligent Information and Database Systems. Springer Berlin Heidelberg, pp. 266-274.

[20] Pham, X. H., Jung, J. J., & Hwang, D. (2012). *Beating Social Pulse: Understanding Information Propagation via Online Social Tagging Systems*. J. UCS, 18(8), 1022-1031.

[21] Jung, J. J. (2012). *Discovering community of lingual practice for matching multilingual tags from folksonomies*. The Computer Journal, 55(3), 337-346.

[22] Nguyen, D. T., Hwang, D., & Jung, J. J. (2014). *Event Detection from Social Data Stream Based on Time-Frequency Analysis*. In Computational Collective Intelligence. Technologies and Applications (pp. 135-144). Springer International Publishing.

[23] Jung, J. J. (2014). *Understanding information propagation on online social tagging systems: a case study on Flickr*. Quality & Quantity, 48(2), 745-754.

[24] Pham, X. H., Jung, J. J., & Le Anh Vu, S. B. P. (2014). *Exploiting social contexts for movie recommendation*. Malaysian Journal of Computer Science, 27(1).

[25] http://en.wikipedia.org/wiki/Wikipedia:Statistics

[26] http://blog.dbpedia.org/2014/09/09/dbpedia-version-2014-released/

[27] http://wiki.linkedmdb.org/Main/Statistics

[28] http://newsroom.fb.com/Key-Facts

[29] https://about.twitter.com/company

[30] http://www.youtube.com/yt/press/statistics.html