

CHUYỂN ĐỔI VÀ PHÂN TÍCH DỮ LIỆU MẠNG XÃ HỘI VỚI MÔ HÌNH DỮ LIỆU RDF

Hoàng Hữu Hạnh*, Phạm Vũ Minh Tú, Vũ Tiến Thành, Ngô Tiến Đạt

Học viện Công nghệ Bưu chính Viễn thông

*Email: hoanghuuhanh@ptit.edu.vn

Ngày nhận bài: 18/11/2019; ngày hoàn thành phản biện: 28/11/2019; ngày duyệt đăng: 20/12/2019

TÓM TẮT

Mạng xã hội ngày càng đóng vai trò rất to lớn và không thể thay thế trong thời đại số hiện nay. Việc sử dụng có hiệu quả dữ liệu mạng xã hội cho các dịch vụ thông tin đa dạng trong phát triển kinh tế, xã hội, khoa học và giáo dục. Dữ liệu mạng xã hội, chẳng hạn như Facebook, Tweeter hay Zalo đều cơ bản được biểu diễn theo mô hình dữ liệu đồ thị với các liên kết định kiểu giữa các đối tượng thông tin. Trong khi đó, RDF và Linked Data là những tiêu chuẩn dữ liệu dựa trên mô hình dữ liệu bộ ba làm cơ sở cho mạng ngữ nghĩa, đã thể hiện tính hiệu quả trong biểu diễn và mô tả tài nguyên dữ liệu trên không gian Web hiện nay. Việc chuyển đổi dữ liệu mạng xã hội trong biểu diễn dạng RDF/Linked Data góp phần tạo nên một nền tảng dữ liệu liên kết có ý nghĩa và sẽ góp phần trong việc nâng cao hiệu quả của việc phân tích dữ liệu của các mạng xã hội.

Từ khóa: Facebook, Linked Data, Mạng xã hội, Phân tích dữ liệu, RDF.

1. MỞ ĐẦU

Mạng xã hội (social network) trở thành một phần quan trọng trong xã hội và nền kinh tế số hiện nay. Tính đa dạng dữ liệu của mạng xã hội và tính dễ kết nối, dễ mở rộng “mạng lưới” thông qua các liên kết thông tin và dữ liệu, tốc độ sản sinh và chia sẻ dữ liệu trên mạng xã hội đã đưa dữ liệu mạng xã hội thành một kho dữ liệu phân tán rất lớn. Nguồn dữ liệu này đã trở thành một môi trường số rất quan trọng tồn tại song song với xã hội thực tế, và cũng ảnh hưởng rất lớn đến những vấn đề của thực tế với cách thức thông tin được chia sẻ, được phân phát cũng như tính đa dạng và mở của nó. Mạng xã hội (MXH) hay dữ liệu mạng xã hội đã trở thành một trong những nguồn thông tin rất năng lực trong các hoạt động của xã hội số hiện nay, từ tin tức, thông tin quảng bá, quảng cáo, truyền thông, v.v... ảnh hưởng đến con người chúng ta trong cách thức giao tiếp và sử dụng nguồn thông tin vô cùng lớn này. Việc phân tích dữ liệu mạng xã hội vì thế sẽ là một vấn đề luôn có tính thời sự và cấp thiết nhằm cung

cấp một thông tin hữu ích, chính xác và dự đoán cũng như dự báo được những trào lưu đang diễn ra trên các mạng xã hội.

Xét về khía cạnh khoa học, dữ liệu mạng xã hội là Dữ liệu Lớn (Big Data) với các tính chất về dung lượng rất lớn (Volume), tính đa dạng của dữ liệu và thông tin (Variety), tốc độ cung cấp dữ liệu trên mạng xã hội (Velocity), tuy nhiên tính chính xác và đáng tin cậy (Veracity) của dữ liệu mạng xã hội là một vấn đề. Nếu xét theo phạm vi nhỏ, khi mỗi người cung cấp thông tin cá nhân của mình, thì chủ quan rằng có sự chính xác; tuy nhiên, hiện nay các dữ liệu được cung cấp trên MXH thường không xuất phát từ các cá nhân sử dụng MXH mà có được từ chia sẻ các nguồn dữ liệu Internet khác. Việc phân tích và phân loại phục vụ lọc thông tin là một vấn đề đang được triển khai ứng dụng, tuy nhiên, tính chất ngữ nghĩa khá bị hạn chế do các tiếp cận trong phân tích dữ liệu hiện nay đều theo các tiếp cận sử dụng các phương pháp truyền thống trong khai phá dữ liệu hay phân tích toán học thông thường. Trong khi đó, Dữ liệu mạng xã hội được biểu diễn thường đi kèm với một lượt đồ hay bộ từ vựng đại diện cho tri thức miền. Do đó phương pháp truyền thống chưa khám phá hết các thông tin liên kết và giàu ngữ nghĩa của dữ liệu MXH hiện nay [1].

Nhằm cung cấp một tiếp cận chuyển đổi và biểu diễn dữ liệu MXH giàu ngữ nghĩa và phục vụ phân tích dữ liệu hiệu quả hơn, bài báo này đề xuất một phương pháp và giới thiệu một mô hình xử lý và phân tích dữ liệu MXH theo tiếp cận biểu diễn và xử lý theo ngữ nghĩa của dữ liệu dựa trên RDF. Bài báo đề cập đến:

- Thu thập dữ liệu MXH từ mạng xã hội Facebook theo cơ chế bán thời gian thực và chuyển đổi sang dạng biểu diễn giàu ngữ nghĩa RDF sử dụng các tri thức hiện có là FOAF [2];
- Phát triển phương pháp xử lý dữ liệu MXH dựa trên hình thức biểu diễn dữ liệu MXH theo mô hình RDF [3].

Với các nội dung như trên, cấu trúc bài báo được chia thành các phần như sau: mô hình dữ liệu RDF được giới thiệu ở phần 2. Phần 3 đề xuất hệ thống chuyển đổi và xử lý dữ liệu MXH nói chung. Và cuối cùng, phát triển hệ thống và kết quả thực nghiệm sẽ được trình bày ở phần 4.

2. MÔ HÌNH DỮ LIỆU RDF CHO DỮ LIỆU MẠNG XÃ HỘI

2.1. Mô hình dữ liệu RDF

Mô hình cơ bản của RDF gồm ba đối tượng sau:

- *Tài nguyên (Resources)*: là tất cả những gì được mô tả bằng biểu thức RDF. Nó có thể là một trang Web, một phần của trang web hay một tập các trang Web.

- *Thuộc tính (Properties)*: thuộc tính, đặc tính, hoặc quan hệ dùng để mô tả tính chất của tài nguyên.
- *Giá trị (Value)*: là giá trị được gán cho thuộc tính.

Mô hình RDF mô tả các tài nguyên thông qua các phát biểu (Statements), mỗi phát biểu là một bộ ba $\langle S, P, O \rangle$, gồm ba thành phần sau :

- *S (Subject - Chủ thể)*: địa chỉ hay vị trí tài nguyên muốn mô tả, được xác định bởi URI.
- *P (Predicate - Vị từ/thuộc tính)*: xác định tính chất của tài nguyên, cũng được xác định bởi URI.
- *O (Object - Đối tượng)*: nội dung gán cho thuộc tính, có thể là tài nguyên hoặc giá trị literal.

• Bộ ba RDF

Mỗi một phát biểu (*subject, predicate, object*) còn gọi là một bộ ba (*triple*).

Ví dụ: Xét phát biểu sau: “**http://www.example.org/index.html** has a **creator** whose value is **John Smith**.”

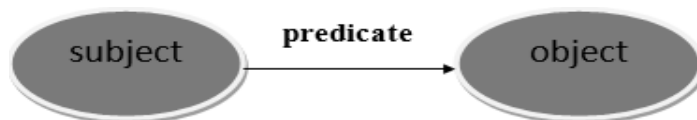
Phát biểu trên được phân ra thành các phần sau:

Subject	<i>http://www.example.org/index.html</i>
Predicate	<i>http://purl.org/dc/elements/1.1/creator</i>
Object	<i>http://www.example.org/staffid/85740</i>

Chúng ta đã biết URI dùng để mô tả bất cứ cái gì trên Web, cho nên ở phát biểu trên thay vì phải dùng từ “*creator*” và “*John Smith*”, ta sẽ dùng URI “*http://purl.org/dc/elements/1.1/creator*” và “*http://www.example.org/staffid/85740*” để mô tả cho từng khái niệm trên.

• Đồ thị RDF

Tập các *triple* hợp lại tạo thành đồ thị RDF (RDF Graph). Các node trong đồ thị có thể là các *subject* và *object* trong *triple* và các cung trong đồ thị là các *predicate*.



Hình 1. Mối quan hệ giữa các thành phần trong bộ ba RDF

2.2. Mô hình RDF dựa trên XML (RDF/XML)

• Cú pháp RDF/XML cơ bản

Mô hình RDF thể hiện một mô hình ở mức trừu tượng để định nghĩa các metadata. Cấu trúc RDF dựa trên cú pháp XML được minh họa trong ví dụ sau.

Ví dụ: Xét phát biểu :

ex:index.html extermis:creation-date "August 16, 1999" .

Một cú pháp RDF/XML để biểu diễn cho phát biểu trên như sau:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:extermis="http://www.example.org/terms/">
  <rdf:Description rdf:about="http://www.example.org/index.html">
    <extermis:creation-date>August 16, 1999</extermis:creation-date>
  </rdf:Description>
</rdf:RDF>
```

- **rdf:resource**

Đối với máy tính, việc sử dụng tên giống nhau chỉ là sự trùng lặp; chẳng hạn Le Huynh B là người dạy môn có mã số CIT3112 có thể không phải là người có ID là 949318 người cũng được gọi tên là Le Huynh B. Điều chúng ta cần là một sự đặc tả hình thức về điều này, ví dụ giảng viên của môn CIT1111 là giảng viên có số mã là 949318, là người có tên là Le Huynh B. Chúng ta có thể đạt được điều này bằng cách sử dụng tham số rdf:resource như sau:

```
<rdf:Description rdf:about="CIT1111">
  <uni:courseName>Discrete Mathematics</uni:courseName>
  <uni:isTaughtBy rdf:resource="949318"/>
</rdf:Description>
<rdf:Description rdf:about="949318">
  <uni:name>Le Huynh B</uni:name>
  <uni:title>Associate Professor</uni:title>
</rdf:Description>
```

- **rdf:type**

Trong các ví dụ, các mô tả tập trung vào hai mục đó là các khoá học (course) và giảng viên (lecturers). Điều này là rõ ràng với con người, nhưng lại không được khai báo một cách hình thức ở bất kỳ đâu, so đó nó không thể được máy tính xử lý được.

Trong RDF, chúng ta có thể tạo ra một phát biểu để giúp làm việc đó bằng cách sử dụng thành phần `rdf:type`.

Ví dụ sau đây là hai mô tả chứa thông tin về kiểu của tài nguyên.

```
<rdf:Description rdf:about="CIT1111">
  <rdf:type rdf:resource="&uni;course"/>
  <uni:courseName>Discrete Mathematics</uni:courseName>
  <uni:isTaughtBy rdf:resource="949318"/>
</rdf:Description>

<rdf:Description rdf:about="949318">
  <rdf:type rdf:resource="&uni;lecturer"/>
  <uni:name>David Billington</uni:name>
  <uni:title>Associate Professor</uni:title>
</rdf:Description>
```

2.3. Truy vấn dữ liệu RDF với SPARQL

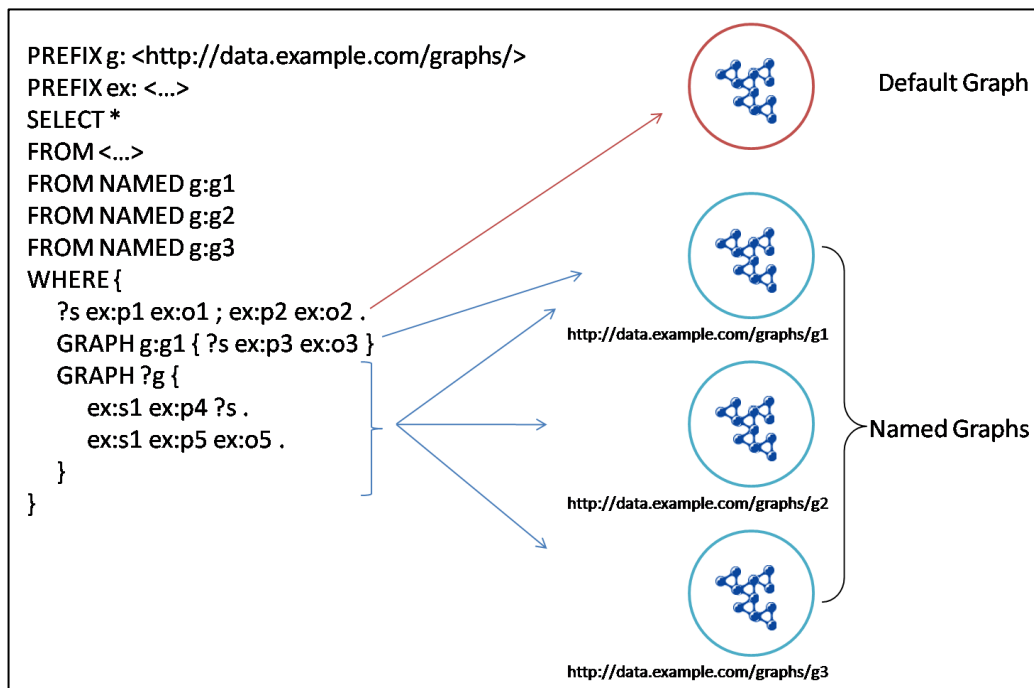
SPARQL – Simple Protocol And RDF Query Language là ngôn ngữ truy vấn và giao thức trao đổi dữ liệu RDF của W3C. SPARQL được phát triển dựa trên các ngôn ngữ RDF trước đây là `rdflib`, `RDQL` và `SeRQL`, và bổ sung một số đặc tính mới [4]. SPARQL có thể được sử dụng để tạo các truy vấn trên các nguồn dữ liệu đa dạng, cho dù dữ liệu được lưu giữ một cách đơn thuần như RDF hay được xem như là RDF thông qua phần mềm trung gian.

Các kho dữ liệu RDF của dữ liệu MXH chứa nhiều các đồ thị RDF và thông tin bản ghi về mỗi đồ thị và nó cho phép một ứng dụng để thực hiện các truy vấn có liên quan đến thông tin từ nhiều hơn một đồ thị. Một Tập dữ liệu RDF (RDF dataset) đại diện cho một tập hợp các đồ thị. Nó bao gồm một đồ thị *mặc định* (*default graph*) và hoặc không có hoặc có nhiều đồ thị được đặt tên (*named graph*). Trong đó các đồ thị được đặt tên được định danh bởi một URI.

Đồ thị được sử dụng cho việc đối sánh một basic mẫu đồ thị cơ sở là *đồ thị đang hoạt động* (*active graph*). Ở trong các phần trước tất cả các truy vấn đều được thực thi đối với một đơn đồ thị điều đó có nghĩa là đồ thị mặc định của RDF Dataset chính là đồ thị đang hoạt động.

Định nghĩa của Tập dữ liệu RDF không hạn chế mối quan hệ của đồ thị được đặt tên và đồ thị mặc định. Thông tin có thể được lặp đi lặp lại trong các đồ thị khác nhau và các mối quan hệ giữa các đồ thị có thể được phơi bày. Có hai chú ý hữu dụng mà ta cần xem xét:

- Để có thông tin trong đồ thị mặc định phải bao gồm các thông tin gốc trên các đồ thị được đặt tên.
- Nên bao gồm các thông tin trên đồ thị đặt tên trong đồ thị mặc định.



Hình 2. Minh họa truy xuất các đồ thị trong RDF dataset

Khi truy vấn một tập hợp các đồ thị, từ khóa GRAPH được sử dụng để đối sánh các mẫu với các đồ thị được đặt tên. GRAPH có thể cung cấp một IRI để lựa chọn một đồ thị hoặc sử dụng một biến chỉ các IRI trong truy vấn RDF Dataset.

Xét ví dụ dưới đây với hai đồ thị với dữ liệu MXH (profile) như sau:

Named graph: <http://example.org/foaf/aliceFoaf>

@prefix foaf: <http://xmlns.com/foaf/0.1/> .

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

_:a foaf:name "Alice" .

_:a foaf:mbox <mailto:alice@work.example> .

_:a foaf:knows _:b .

_:b foaf:name "Bob" .

_:b foaf:mbox <mailto:bob@work.example> .

_:b foaf:nick "Bobby" .

```
_:b rdfs:seeAlso <http://example.org/foaf/bobFoaf> .
<http://example.org/foaf/bobFoaf>
  rdf:type    foaf:PersonalProfileDocument .

# Named graph: http://example.org/foaf/bobFoaf
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
_:z foaf:mbox <mailto:bob@work.example> .
_:z rdfs:seeAlso <http://example.org/foaf/bobFoaf> .
_:z foaf:nick  "Robert" .
<http://example.org/foaf/bobFoaf>
  rdf:type    foaf:PersonalProfileDocument .
```

Sử dụng truy vấn SPARQL chúng ta có thể truy vấn các thông tin “lệch” nhau được lưu trữ trên các profile được cho ở trên:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?src ?bobNick
FROM NAMED <http://example.org/foaf/aliceFoaf>
FROM NAMED <http://example.org/foaf/bobFoaf>
WHERE {
  GRAPH ?src {
    ?x foaf:mbox <mailto:bob@work.example> .
    ?x foaf:nick ?bobNick }
}
```

Sẽ cho ra kết quả là:

src	bobNick
http://example.org/foaf/aliceFoaf	"Bobby"
http://example.org/foaf/bobFoaf	"Robert"

Xem xét một ví dụ khác, ta xét một truy vấn khác dựa trên các tập dữ liệu MXH dạng RDF như đã cho ở trên:

Chuyển đổi và phân tích dữ liệu mạng xã hội với mô hình dữ liệu RDF

```
PREFIX data: <http://example.org/foaf/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?mbox ?nick ?ppd
FROM NAMED <http://example.org/foaf/aliceFoaf>
FROM NAMED <http://example.org/foaf/bobFoaf>
WHERE {
  GRAPH data:aliceFoaf
  { ?alice foaf:mbox <mailto:alice@work.example> ;
    foaf:knows ?whom .
    ?whom foaf:mbox ?mbox ;
    rdfs:seeAlso ?ppd .
    ?ppd a foaf:PersonalProfileDocument .
  } .
  GRAPH ?ppd
  { ?w foaf:mbox ?mbox ;
    foaf:nick ?nick
  }
}
```

Kết quả truy vấn:

mbox	nick	ppd
<mailto:bob@work.example>	"Robert"	<http://example.org/foaf/bobFoaf>

Kết quả này cho thấy được cách thức truy vấn dữ liệu MXH dạng RDF được thể hiện bằng các truy vấn SPQARQL thuận lợi và cách để kiểm chứng thông tin.

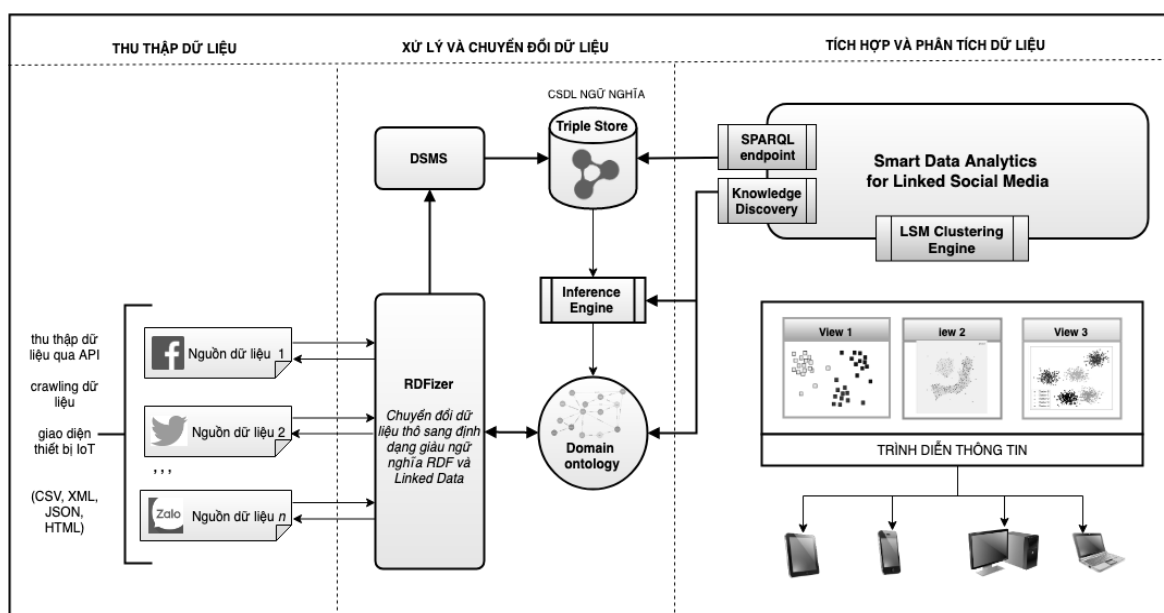
3. KIẾN TRÚC HỆ THỐNG KHUNG XỬ LÝ

Hệ thống Chuyển đổi và Phân tích dữ liệu mạng xã hội (gọi tên là **iSDA**) được thiết kế với mục tiêu là thu thập các nguồn dữ liệu mạng xã hội, sau đó chuyển đổi biểu diễn dạng RDF và ứng dụng các phương pháp phân tích dữ liệu MXH. Hệ thống iSDA vì vậy sẽ nhắm đến các tác vụ sau:

- Thu thập dữ liệu MXH theo cơ chế bán thời gian thực;
- Xử lý và chuyển đổi dữ liệu sang dạng biểu diễn giàu ngữ nghĩa;
- Tích hợp và Phân tích dữ liệu MXH theo mô hình RDF.

3.1. Module cho Thu thập dữ liệu Mạng xã hội

Dữ liệu mạng xã hội sẽ được thu thập thông qua các crawler được lập trình để quét những nhóm người sử dụng, fanpage, các kênh theo đặc thù. MXH phổ biến nhất hiện nay tại Việt Nam là Facebook thì không cung cấp các API để có thể giao tiếp thu thập thông tin dễ dàng, do đó, đề tài sẽ đề xuất mô hình crawler (social media crawler) phù hợp để quét những dữ liệu MXH có trọng tâm theo cơ chế bán thời gian thực. Dữ liệu Facebook cơ bản là dạng dữ liệu đồ thị (graph data), do đó việc chuyển đổi qua mô hình dữ liệu RDF là một sự chuyển đổi tự nhiên và có thể giữ được các mối quan hệ mang thông tin và dữ liệu của đặc trưng dữ liệu MXH. Một cơ chế crawling dữ liệu đồ thị là hướng tiếp cận của đề tài đề xuất.



Hình 3. Kiến trúc hệ thống chuyển đổi và xử lý dữ liệu dựa trên công nghệ Web ngữ nghĩa.

3.2. Module cho Xử lý và Chuyển đổi dữ liệu MXH

Như vừa đề cập ở trên, dữ liệu MXH cơ bản là dạng đồ thị, và các liên kết giữa các đối tượng dữ liệu của dữ liệu MXH là các liên kết đối tượng với đối tượng, tức là các mối quan hệ có chứa thông tin và dữ liệu mang ngữ nghĩa. Việc biểu diễn các dữ liệu MXH Facebook nếu dùng cơ chế mô hình và biểu diễn truyền thống sẽ làm mất đi các ngữ nghĩa hiện có của dữ liệu MXH mà ta thu thập được.

Việc xử lý sơ bộ các dữ liệu thu thập được và đảm bảo tính ngữ nghĩa của dữ liệu cũng như các mối quan hệ, đề tài đề xuất xây dựng một mô hình biểu diễn dữ liệu và thông tin của MXH thu thập được gồm 2 phần:

- Các ontology miền: đóng vai trò như cơ sở tri thức biểu diễn không gian các khái niệm hiện có trong các MXH cũng như sử dụng lại và liên đến đến các không gian và cơ sở tri thức đã được định nghĩa trên mạng dữ liệu liên kết toàn cầu (Linked Open Data cloud). Một phần quan trọng của cơ sở tri thức này là một Máy suy diễn dựa trên tri thức miền này.
- Dữ liệu MXH được thu thập sẽ được xử lý (làm sạch) và chuyển đổi sang dạng RDF/Linked Data đảm bảo được ngữ nghĩa của dữ liệu cũng như bảo toàn các mối quan hệ thông tin và đối tượng là đặc tính quan trọng của dữ liệu MXH. Mô hình dữ liệu của RDF/Linked Data/OWL là mô hình dữ liệu *bộ ba (triple)* bao gồm $\langle S, P, O \rangle$; trong đó S là Subject chứa URI của một tài nguyên (đối tượng dữ liệu), P là thuộc tính (Property) của S, biểu diễn ngữ nghĩa của S và có giá trị là O (object) là giá trị Literal hoặc là một tài nguyên (đối tượng dữ liệu) khác.

Ví dụ:

$\langle \text{facebook.com/hhhanh} \rangle \langle \text{có_tên} \rangle \text{“Hanh Hoang”}$

$\langle \text{facebook.com/hhhanh} \rangle \langle \text{là_bạn} \rangle \langle \text{facebook.com/pvmtu} \rangle$

$\langle \text{facebook.com/pvmtu} \rangle \langle \text{có_status} \rangle \langle \text{facebook.com/pvmtu/post/111222333} \rangle$

$\langle \text{facebook.com/pvmtu/post/111222333} \rangle \langle \text{biểu_cảm} \rangle \text{“Happy”}$

3.3. Module cho Tích hợp và Phân tích dữ liệu MXH

Dữ liệu sau khi được chuyển đổi (và lưu trữ trong CSDL ngữ nghĩa Triple Store) sẽ được xử lý tích hợp theo hình thức mashup: ánh xạ các thuộc tính ngữ nghĩa từ ontology miền về miền tri thức chính theo ngữ cảnh để xử lý. Việc truy xuất các CSTT ontology có thể sử dụng các cơ chế suy diễn để tìm ra các mối quan hệ thông tin ẩn chứa trong các dữ liệu MXH;

Việc biểu diễn dữ liệu MXH theo mô hình RDF/Linked Data cho phép biểu diễn giàu hơn các ngữ nghĩa của dữ liệu và thông tin thông qua các thuộc tính và các khái niệm tri thức miền. Các thuật toán về phân tích dữ liệu cho các dữ liệu dạng RDF/Linked Data sẽ được phát triển để tận dụng tối đa hình thức biểu diễn này, cũng như qua các tri thức miền ontology. Các thuật toán về phân tích dữ liệu MXH theo hình thức và mô hình biểu diễn giàu ngữ nghĩa này cũng sẽ tập trung vào: việc phân cụm dữ liệu và dự báo về xu hướng của các luồng thông tin MXH.

4. iSDA - HỆ THỐNG PHÂN TÍCH DỮ LIỆU MẠNG XÃ HỘI FACEBOOK

4.1. Thu thập dữ liệu

Dữ liệu mạng xã hội đóng một vai trò quan trọng trong kỷ nguyên của cuộc cách mạng công nghiệp 4.0 hiện nay. Dữ liệu mạng xã hội hay các nguồn dữ liệu cơ bản được biểu diễn dưới dạng dữ liệu đồ thị. Việc sử dụng RDF để chuyển đổi dữ liệu mạng xã hội đối với báo chí dữ liệu là một tiếp cận tự nhiên.

Các nguồn dữ liệu được thu thập từ Facebook với hơn 88 Triệu tài khoản Facebook; 3 Triệu groups; 2 Triệu pages; 30 Triệu tín hiệu/ngày. Các dạng dữ liệu của Facebook: Posts/Pictures/Videos, Likes/Shares/Comments.

- Hệ thống iSDA thu thập:
 - o Blog/news/forum: 900 đầu báo chính thống, 100 trang tổng hợp tin tức, 88 Forum, 150 Blog
 - o Các loại dữ liệu thu thập được: Bài báo, Bài viết/Hình ảnh/Videos, Bình luận, cảm xúc,...
- Tổng số lượng thu thập:
 - o 30 triệu bài viết / ngày từ các nguồn MXH, Báo chí, Forums, Blog.
 - o Dễ dàng bổ sung nguồn dữ liệu , tài khoản FB, trang báo chí, forum, blog cần theo dõi , thu thập dữ liệu
 - o Lưu trữ dữ liệu trực tiếp trên hệ thống , có thể xem đ ư ợ c nội dung ngay cả khi bài viết bị xóa

Do đó phát triển hệ thống phân tích dữ liệu mạng xã hội nhằm cung cấp một giải pháp tổng thể về thu thập và phân tích từ nhiều nguồn dữ liệu khác nhau như Mạng xã hội, báo chí,... Cho phép theo dõi các nhân vật chủ đề sự kiện cần lưu ý, tự động cảnh báo khi phát hiện vấn đề hỗ trợ phân tích cảnh báo sớm khủng hoảng truyền thông.

4.2. Các thành phần hệ thống

Kiến trúc hệ thống Phân tích dữ liệu Mạng xã hội bao gồm các module sau:

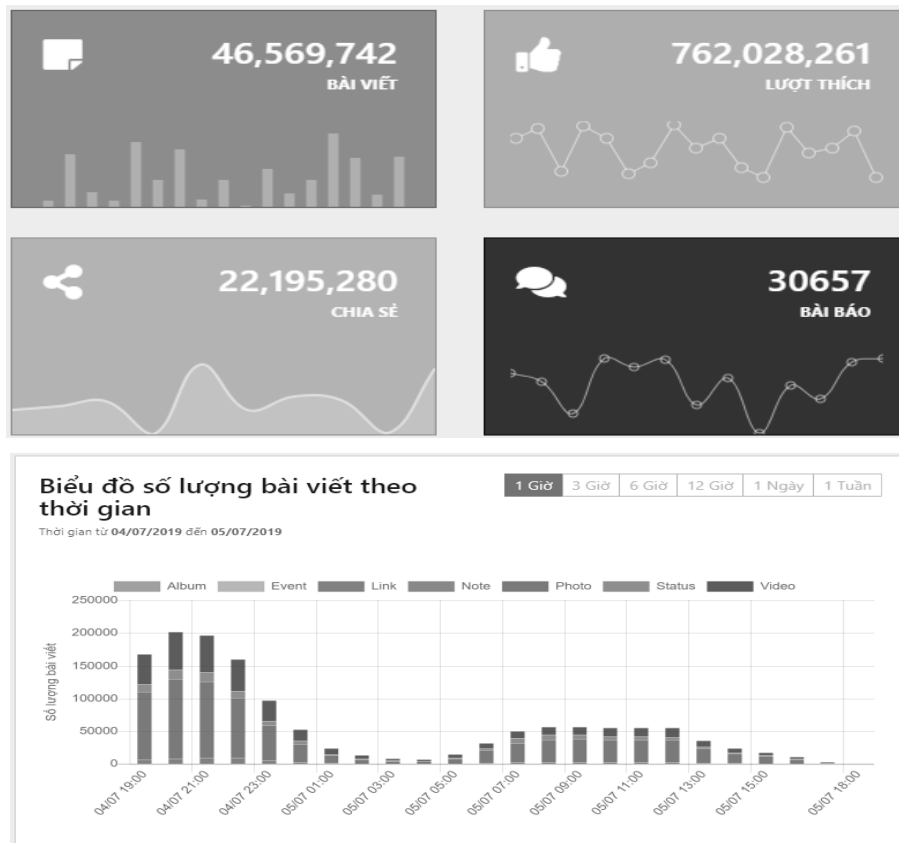
- Thu thập dữ liệu: thu thập dữ liệu từ các nguồn dữ liệu xác định và chuyển đổi qua mô hình RDF/Linked Data [5][7].
- Khai phá và Phân tích dữ liệu: Sử dụng SPARQL và các phương pháp cơ bản để tổng hợp phân tích dữ liệu [4][6].

- Lưu trữ và truy vấn: lưu trữ mô hình dữ liệu vật lý và các hệ quản trị cơ sở dữ liệu cho mô hình bộ ba RDF và Linked Data [8].
- Và cung cấp các ứng dụng: cung cấp các khung nhìn thông tin được trích xuất dữ liệu.

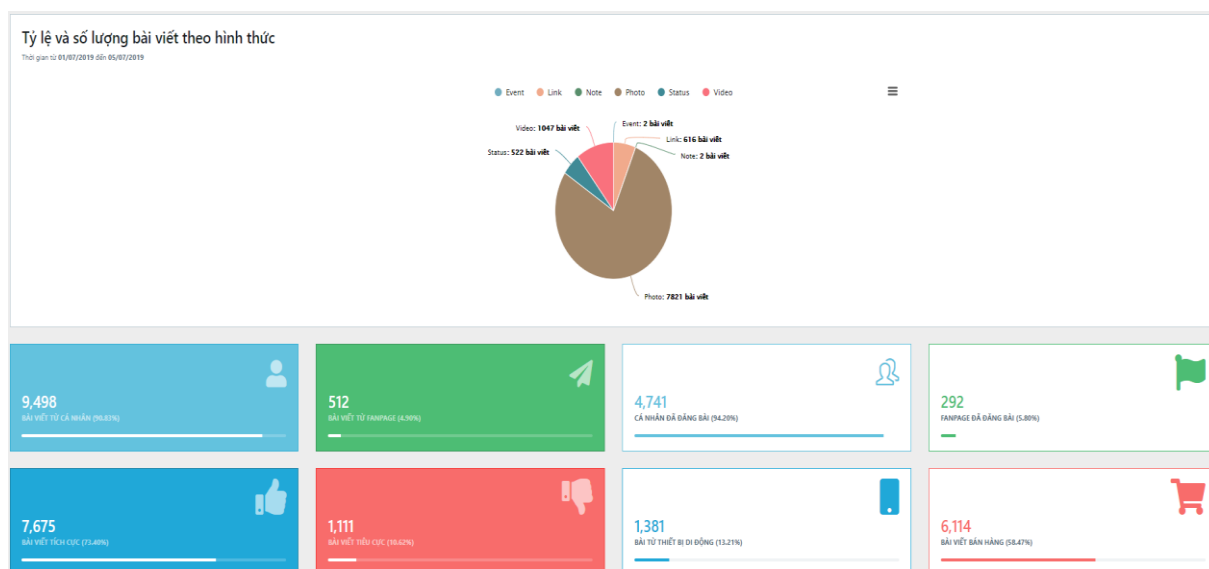
4.3. Các tính năng của hệ thống

Hệ thống iSDA được host tại địa chỉ <http://222.255.146.19:8888>, cung cấp các tính năng sau:

- Phân tích xu hướng chủ đề sự kiện
- Thu thập và cập nhập thông tin từ mạng xã hội , báo chí , forums và blogs
- Quản lý, thống kê nguồn thông tin
- Theo dõi, cảnh báo thông tin



Hình 4. Dashboard hiển thị thông tin và dữ liệu thống hợp với biểu đồ số lượng thống kê bài viết Facebook.



Hình 5. Kết quả hệ thống tự động thu thập và theo dõi biểu tình tại Hongkong từ 01/07/2019 đến 05/07/2019.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo trình bày kết quả nghiên cứu ứng dụng trong mô hình hoá dữ liệu với thông tin ngữ nghĩa. Mô hình RDF/Linked Data đáp ứng các yêu cầu cho việc biểu diễn và phân tích dữ liệu với các trích dẫn và chú thích ngữ nghĩa như dữ liệu mạng xã hội. Việc triển khai hệ thống iSDA vào thực tiễn là một đóng góp thực tiễn của bài báo. Trong tương lai, nhóm tác giả sẽ tập trung đẩy mạnh nghiên cứu các phương pháp phân tích dữ liệu cải tiến để tận dụng lợi thế của việc biểu diễn theo ngữ nghĩa này.

TÀI LIỆU THAM KHẢO

- [1]. Paolillo, J. C., & Wright, E. (2006). Social network analysis on the semantic web: Techniques and challenges for visualizing FOAF. In *Visualizing the semantic web* (pp. 229-241).
- [2]. FOAF ontology (2014). Website: <http://xmlns.com/foaf/spec/>
- [3]. Resource Description Framework (RDF) (2014). Website: <https://www.w3.org/RDF/>
- [4]. SPARQL 1.1 Query Language (2013). Website: <https://www.w3.org/TR/sparql11-query/>
- [5]. M. San Martín and C. Gutierrez (2009). Representing, Querying and Transforming Social Networks with RDF/SPARQL, *Proc. ESWC*. Boston, LNCS 5554, pp. 293-307.
- [6]. S. Elbassuoni, M. Ramanath, R. Schenkel, G. Weikum (2010). Searching RDF Graphs with SPARQL and Keywords, *IEEE Data Eng. Bull.*, 33(1), pp.16-24.
- [7]. Linked Data (2019). Website: <https://www.w3.org/standards/semanticweb/data>
- [8]. Apache Jena (2019). Website: <https://jena.apache.org/>

SOCIAL NETWORK DATA TRANSFORMATION AND ANALYSIS USING RDF DATA MODEL

Hoang Huu Hanh*, Pham Vu Minh Tu, Vu Tien Thanh, Ngo Tien Dat

Posts and Telecommunications Institute of Technology

*Email: hoanghuuhanh@ptit.edu.vn

ABSTRACT

Social networks play important roles in our present digital age. Effectively using social networks' data can contribute into the eco-socio development through its diversity of information in economic, social and educational areas. The social networks' data such as Facebook, Tweeter or Zalo are basically presented in the form of graph data model with links between information objects or data items. Meanwhile, RDF or Linked Data are standards for data modelling in triple model which forming a foundation for semantic networks and showing its effectiveness in resource representation in our Web data space. Therefore, transforming social network data into the representation of RDF/Linked Data eventually makes a meaningful linked data foundation in advancing social data analysis.

Keywords: Facebook, Linked Data, Social network, Data analysis, RDF.



Hoàng Hữu Hạnh sinh ngày 13/04/1974 tại Huế. Năm 1996, ông tốt nghiệp Cử nhân ngành Toán-Tin học tại Trường Đại học Sư phạm Huế, Thạc sĩ khoa học ngành Công nghệ thông tin tại Trường Đại học Bách khoa Hà Nội. Năm 2007, ông nhận học vị Tiến sĩ chuyên ngành Hệ thống thông tin tại Trường Đại học Công nghệ Vienna, Cộng hoà Áo. Năm 2012, ông nhận Chức danh Phó giáo sư tại Việt Nam. Từ năm 1996-2018 đến nay, ông là Giảng viên Khoa Công nghệ Thông tin, Trường ĐH Khoa học Huế; đồng thời đảm trách các nhiệm vụ quản lý tại Đại học Huế từ 2008 đến 2018. Từ tháng 10 năm 2018 ông là Giảng viên cao cấp tại Học viện Công nghệ Bưu chính Viễn thông (PTIT). Hiện nay là Giám đốc Trung tâm Đào tạo Quốc tế của PTIT.

Lĩnh vực nghiên cứu: Biểu diễn tri thức, Web ngữ nghĩa, Linked Data, Ontology, Công nghệ phần mềm, Công nghệ dữ liệu, Quản lý quy trình nghiệp vụ.



Phạm Vũ Minh Tú sinh ngày 30/09/1990 tại thành phố Hà Nội. Năm 2013, ông tốt nghiệp kỹ sư ngành Kỹ thuật Điện – Điện tử tại Học Viện Công Nghệ Bưu Chính Viễn Thông. Năm 2015, ông tốt nghiệp thạc sĩ chuyên ngành Kỹ thuật Viễn Thông tại trường Học Viện Công Nghệ Bưu Chính Viễn Thông. Từ năm 2015 đến nay, ông giảng dạy tại trường Học Viện Công Nghệ Bưu Chính Viễn Thông.

Lĩnh vực nghiên cứu: Truyền thông quang không dây, Xử lý tín hiệu số, Dữ liệu đa phương tiện



Vũ Tiến Thành sinh ngày 02/01/1991 tại Thành phố Hà Nội. Năm 2013, ông tốt nghiệp Cử nhân ngành Báo chí và Truyền thông tại Trường Đại học Khoa học Xã hội & Nhân Văn, Đại học QGHN. Năm 2016, ông tốt nghiệp thạc sĩ chuyên ngành Báo chí và Truyền thông tại Trường Đại học Khoa học Xã hội & Nhân Văn, Đại học QGHN. Từ năm 2016 đến nay giảng dạy tại Học viện Công nghệ Bưu chính Viễn Thông.

Lĩnh vực nghiên cứu: Báo chí và Truyền thông Mạng.



Ngô Tiến Đạt sinh ngày 01/07/1998 tại Thành phố Bắc Ninh, là sinh viên năm cuối Chương trình Kỹ sư ngành Công nghệ Đa Phương Tiện tại Học Viện Công Nghệ Bưu Chính Viễn Thông.

Lĩnh vực nghiên cứu: Truyền thông mạng, Dữ liệu đa phương tiện

