

BIỂU DIỄN NGŨ NGHĨA TRONG WEB NGŨ NGHĨA

Th.S. Trần Vĩnh Xuyên

Khoa Công nghệ Thông tin, Trường Cao đẳng Kỹ thuật Lý Tự Trọng

Email: xuyen_dhbk@yahoo.com.au

Tóm tắt

Với sự phát triển của công nghệ thông tin đặc biệt là sự phát triển về cơ sở hạ tầng lẫn công nghệ trên Internet, đã làm thay đổi cách thức làm việc của con người. Khi chúng ta biểu diễn được ngữ nghĩa thông tin của tài nguyên trên Web, thì chúng ta đã giúp cho máy tính có khả năng hiểu dữ liệu mà chúng ta mô tả gần tương tự như chính chúng ta hiểu về vấn đề nào đó. Từ ấy máy tính có thể hỗ trợ thay thế chúng ta tìm kiếm dữ liệu này một cách hiệu quả và chính xác hơn về ngữ nghĩa mà ta mong muốn. Trong bài viết này tôi xin trình bày cách biểu diễn ngữ nghĩa trong Web ngữ nghĩa.

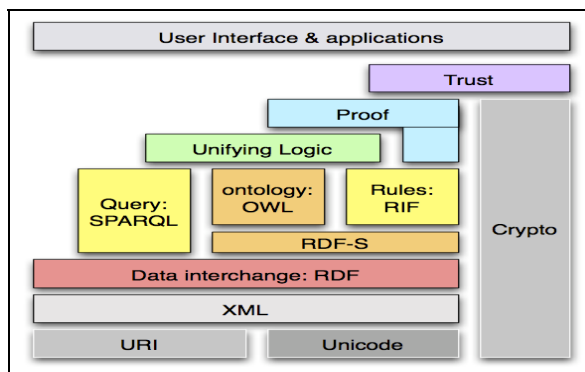
1. TỔNG QUAN WEB NGŨ NGHĨA (WEBNN)

1.1. Web ngữ nghĩa

Theo định nghĩa của Tổ chức World Wide Web (W3C) thì WebNN được hiểu như sau:

“WebNN là một cách nhìn về cách tổ chức dữ liệu: đó là ý tưởng về việc dữ liệu trên Web được định nghĩa và liên kết theo một cách, mà nó có thể được sử dụng bởi máy tính với mục đích không chỉ cho việc hiển thị, mà còn tự động hóa, tích hợp và sử dụng lại dữ liệu qua các ứng dụng khác nhau”.

Kiến trúc phân tầng của WebNN [3] được mô tả trong hình sau. Trong đó, các tầng trên kế thừa các tầng thấp hơn với cơ sở là các chuẩn để mô tả siêu dữ liệu.



Hình 1.1: Kiến trúc phân tầng của WebNN

Kiến trúc phân tầng của web ngữ nghĩa ngày nay, ta thấy có bảy tầng kiến trúc. Trong đó, với hệ thống Web hiện tại (World Wide Web) là đang ở tầng thứ hai.

1.2. Ontology

Thuật ngữ Ontology [2] bắt nguồn từ triết học, nó được sử dụng như tên của một lĩnh vực nghiên cứu về sự tồn tại của tự nhiên, xác định các vật thể trong tự nhiên và làm thế nào mô tả chúng. Chẳng hạn như quan sát thế giới thực, xác định các đối tượng và sau đó nhóm chúng lại thành các lớp trừu tượng dựa trên thuộc tính chung.

Tuy nhiên, trong những năm gần đây, Ontology đã trở thành một thuật ngữ được biết đến nhiều trong lĩnh vực khoa học máy tính và có ý nghĩa xa so với nghĩa ban đầu của nó,

Ontology được xem như là “linh hồn” của Web 3.0. Chúng giúp con người và máy có thể hợp tác, cùng với nhau làm việc, giúp máy có thể “hiểu” và có khả năng xử lý thông tin hiệu quả. Các Ontology được phát triển trong lĩnh vực trí tuệ nhân tạo để sử dụng lại và chia sẻ tri thức được thuận tiện hơn. Đầu những năm 1990, Ontology trở thành một chủ đề nghiên cứu phổ biến. Ontology được nghiên cứu bởi một số cộng đồng nghiên cứu trí tuệ nhân tạo, bao gồm kỹ sư tri thức, xử lý ngôn ngữ tự nhiên và biểu diễn tri thức.

Một định nghĩa chung cho Ontology là: Ontology là một đặc tả hình thức của sự khái niệm hóa về một lĩnh vực ứng dụng cụ thể. Định nghĩa này nhấn mạnh hai điểm chính: sự khái niệm hóa là hình thức và do đó cho phép suy diễn bởi máy tính, và một ontology trên thực tế được thiết kế cho một miền ứng dụng cụ thể nào đó. Các ontology bao gồm các khái niệm, các quan hệ, các thể hiện và các tiên đề.

1.3. Khung mô tả tài nguyên

RDF (Resource Description Framework) là một “bộ khung” được sử dụng để mô tả các nguồn tài nguyên trên Internet. RDF cung cấp một mô hình dữ liệu, và một cú pháp đơn giản sao cho các hệ thống độc lập có thể trao đổi và sử dụng được nó. RDF được thiết kế sao cho hệ thống máy tính có thể hiểu được và có thể đọc được thông tin, chứ không phải để trình bày dữ liệu cho người dùng.

Cấu trúc căn bản của một phát biểu RDF rất đơn giản. Gồm 3 phần:

- Subject: là cái mà ta đề cập đến, được xác định bởi URI.
- Predicate: là thuộc tính của chủ thể có kiểu metadata được xác định bởi URI hay bnode.
- Object: là giá trị của thuộc tính được xác định bởi URI, Bnode, Literal.

Ví dụ:



Hình 1.2: Khung mô tả tài nguyên RDF

2. BIỂU DIỄN NGỮ NGHĨA TRONG WEB NGỮ NGHĨA

2.1. RDFa

Với RDF thì chúng ta đã thành công trong việc biểu diễn dữ liệu từ thế giới thực vào máy tính. Ắt hẳn làm sao cho máy tính có thể hiểu được thông tin từ thế giới thực tương tự như cách hiểu của chúng ta (chúng ta đang xét về ngữ nghĩa), đây chính là một vấn đề mà RDF chưa giải quyết được. Vì vậy sự ra đời của RDFa [5,6] là sự cần thiết và hợp lý.

Web ngày nay được tạo ra không chỉ yêu cầu thể hiện được thông tin, mối quan hệ giữa các thông tin trong nội bộ của một trang web mà còn đòi hỏi sự liên kết thông tin giữa các trang với nhau. Từ đó giúp cho việc tìm kiếm thông tin trên web trở nên thật dễ dàng và chính xác về nội dung và ngữ nghĩa. Với RDF thì chúng ta chỉ dừng lại ở chỗ mô tả chính xác được dữ liệu một cách đơn giản, chưa mô tả được một cách trọn vẹn thế giới thực vào máy tính, và đặc biệt RDF chưa thể thể hiện được trạng thái của các văn bản được nhìn thấy ở hiện tại và các liên kết với các dữ liệu mà máy có thể đọc mà không trùng lặp về nội dung.

RDFa ra đời nhằm bổ sung vào RDF để giải quyết các vấn đề mà thế giới thực yêu cầu.

Ví dụ:

- Chúng ta đã tạo ra một trang Web như sau:



Hình 2.1: Sự hiểu không tương đồng giữa máy tính và con người

Hình bên trái thể hiện sự hiểu của máy tính, hình bên phải thể hiện sự hiểu của con người trên cùng một tài liệu.

Rõ ràng với đến thời điểm này thì giữa chúng ta và máy tính đã không có tiếng nói chung. Chúng ta đã hiểu cấu trúc dữ liệu của một tài liệu thiên về ngữ nghĩa của tài liệu, còn máy tính thì hiểu tài liệu theo cách cứng nhắc (không xét gì về ngữ nghĩa của tài liệu). Đây chính là mấu chốt mà dẫn đến bất đồng giữa máy tính và con người không thể có tiếng nói chung.

Để giải quyết vấn đề này chúng ta có thể bổ sung một số ngữ nghĩa vào trong tài liệu.

2.2. Một số chuẩn biểu diễn ngữ nghĩa

2.2.1. FOAF (Friend Of A Friend)

FOAF [3] là bộ từ vựng đã được xây dựng để mô tả thông tin về con người (people), nhóm (group), tổ chức (organization) và các thông tin liên quan.

Ví dụ để lưu thông tin của một người ta cần có những thông tin sau: tên, tuổi, nghề nghiệp, giới tính, ngày tháng năm sinh, hình ảnh, có những tài liệu nào.v.v. thì trong FOAF đã cung cấp sẵn cho chúng ta một bộ từ vựng cơ bản để có thể mô tả những thông tin này.

Xét một ví dụ sử dụng FOAF

```
<foaf:Person>
  <foaf:name>Xuyen</foaf:name>
  <foaf:mbox rdf:resource=mailto:xuyen@yahoo.com/>
  <foaf:homepage rdf:resource="http://rdfweb.org/people/xuyen/" />
  <foaf:img rdf:resource="http://rdfweb.org/people/danbri/mugshot/xuyen.jpeg" />
</foaf:Person>
```

Tài liệu trên mô tả một người (foaf:Person) có tên (foaf:name) là 'Xuyen' và có địa chỉ mail (foaf:mbox) là xuyen@yahoo.com, có một trang web (foaf:homepage) homepage là http://rdfweb.org/people/xuyen/ và có một hình ảnh (foaf:image) lưu tại địa chỉ http://rdfweb.org/people/danbri/mugshot/xuyen.jpeg.

2.2.2. Giới thiệu chuẩn Dublin Core

Chuẩn Dublin Core là chuẩn dùng để mô tả dữ liệu trong các Metadata nhằm khai thác các tài liệu trên các web site thông qua mạng Internet. Chuẩn Dublin Core bao gồm 15 yếu tố được công bố và thống nhất từ các cuộc hội thảo mang tầm cỡ quốc tế và mang ý nghĩa kết hợp của các ngành khoa học: thư viện, tin học, bảo tàng, mã hoá văn bản và các lĩnh vực khác có liên quan.

NỘI DUNG	SỞ HỮU TRÍ TUỆ	THUYẾT MINH
Ảnhan đề	Tác giả	Ản gày tháng
Đề mục	Tác giả phụ	Mô tả vật lý
Mô tả	Xuất bản	Định danh
Loại hình	Bản quyền	Ản gôn ngữ
Ản guồn gốc		
Liên kết		
Ản ời chứa		

Bảng 2.2.2: Bảng phân loại các yếu tố của Dublin Core

Các yếu tố cơ bản của Dublin Core đều mang thuộc tính lựa chọn và có thể lặp lại. Mỗi yếu tố cũng có một giới hạn những hạn định, thuộc tính nhằm diễn giải chính xác ý nghĩa của các yếu tố.

1. Ẩnhan đề (Title): Tên của nguồn thông tin thường do tác giả hoặc nhà xuất bản đặt cho tài liệu.

2. Tác giả (Creator): Ẩn gười hoặc cơ quan chịu trách nhiệm chính về nội dung trí tuệ của nguồn thông tin

3. Đề mục (Subject): Chủ đề của nguồn thông tin và được thể hiện bằng từ vựng có kiểm soát gồm tiêu đề đề mục, số phân loại,...

4. Mô tả (Description): Phần thể hiện nội dung của nguồn thông tin bao gồm cả phần tóm tắt của tư liệu văn bản hoặc nội dung của tư liệu nghe nhìn

5. Xuất bản (Publisher): Cơ quan, tổ chức chịu trách nhiệm tạo lập, xuất bản nguồn thông tin trong định dạng thực.

6. Tác giả phụ (Contributor): Cá nhân hay tổ chức có những đóng góp về mặt trí tuệ cho tư liệu nhưng không phải là tác giả chính.

7. Ẩn gày tháng (Date): ngày tháng có liên quan đến việc tạo lập, xuất bản hay công bố tư liệu.

8. Loại hình (Type): hình thức vật chứa nội dung tư liệu

9. Mô tả vật lý (Format): Định dạng vật lý và kích thước của tư liệu như kích cỡ, thời lượng,.. Định dạng cũng còn được dùng để chỉ rõ phần mềm và phần cứng cần thiết để sử dụng tư liệu.

10. Định danh tư liệu (Identifier): Là một dãy ký tự hoặc số nhằm thể hiện tính đơn nhất của tư liệu như: URLs và URẢ s, ISBẢ , ISSẢ ,...

11. Ẩn guồn gốc (Source): Ẩn guồn gốc mà tư liệu được tạo thành, yếu tố này có thể bao gồm siêu dữ liệu về nguồn thông tin thứ hai nhằm khai thác tư liệu hiện hành.

12. ả ngôn ngữ (Language): ả ngôn ngữ của nội dung tư liệu, được thành lập theo quy tắc RFC 1766.

13. Liên kết (Relation): Một định danh cho nguồn thứ hai và những mối quan hệ của nó với tư liệu hiện hành. Yếu tố này thể hiện những kết nối giữa những nguồn tư liệu có liên quan.

14. ả ời chứa (Coverage): ả hững đặc tính về không gian và/hoặc thời gian của tư liệu. Không gian nơi chứa chỉ ra một vùng sử dụng địa danh hoặc toạ độ. Đặc tính thời gian trong yếu tố này chỉ ra khoảng thời gian mà tư liệu đề cập tới và thường sử dụng tên thời kỳ như thời kỳ Đồ đá.

15. Bản quyền (Rights): Thông tin về tình trạng bản quyền, kết nối tới thông tin về tình trạng bản quyền hoặc dịch vụ cung cấp thông tin bản quyền cho tư liệu.

Các yếu tố mở rộng:

Thực tế sử dụng Dublin Core cho thấy mỗi yếu tố cơ bản còn gộp chứa trong nó một vài thành tố phụ nhằm diễn đạt chi tiết hơn nội dung chính yếu tố đó. Các thành tố phụ được coi là các yếu tố mở rộng và được thể hiện thông qua những khung mã hoá cụ thể.

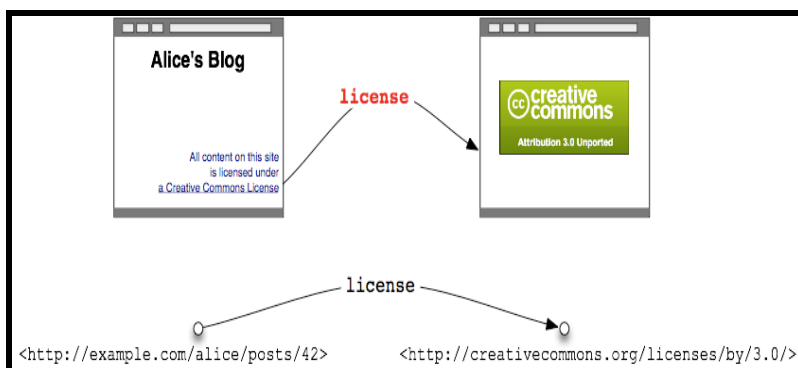
Ví dụ:

- Xác định bản quyền của tài nguyên:

Trong quá khứ chúng ta đã sử dụng như sau:

```
...
Sở hữu bản quyền website này tại
<a href="http://creativecommons.org/licenses/by/3.0/">
    a Creative Commons License
</a>
```

ả ếu viết như trên thì rõ ràng Browser (trình duyệt) sẽ không thể nào hiểu nổi đây chính là mối liên kết để xác định bản quyền của tài liệu.



Hình 2.2: Kết nối thể hiện bản quyền của trang hai web.

ả hưng nếu như chúng ta có thêm một cải tiến nhỏ đó là thêm một thuộc tính `rel="license"` vào trong tag thì lúc ấy Browser sẽ có thể hiểu được ý nghĩa của đoạn tài liệu này đúng như chúng ta hiểu.

```
...
Sở hữu bản quyền website này tại
<a rel="license" href="http://creativecommons.org/licenses/by/3.0/">
```

a Creative Commons License

.

- ả hữ Title và Author

Trong quá khứ chúng ta đã sử dụng như sau:

```
<div>
  <h2>Tiêu Đề Trang Web</h2>
  <h3>ả gười tạo ra website này là Alice</h3>
  ...
</div>
```

ả hưng nếu như chúng ta có thêm một cải tiến nhỏ đó là thêm một thuộc tính `property="dc:title"` và `property="dc:creator"` vào trong tag thì lúc ấy Browser sẽ có thể hiểu được ý nghĩa của đoạn tài liệu này đúng như chúng ta hiểu.

```
<div xmlns:dc="http://purl.org/dc/elements/1.1/">
  <h2 property="dc:title"> Tiêu Đề Trang Web </h2>
  <h3 property="dc:creator">
    ả gười tạo ra website này là Trần Vĩnh Xuyên
  </h3>
  ...
</div>
```

- o Ở đây `property="dc:title"` : Browser hiểu chính là tiêu đề (title).
- o Ở đây `property="dc:creator"` : Browser hiểu chính là người tạo ra (creator).

Kết Luận Và Hướng Phát Triển

Kết luận

Thể hiện được ngữ nghĩa trong web ngữ nghĩa ta cần chúng ta dựa vào hai vấn đề chính đó là:

- o Dùng RDF để mô tả dữ liệu (tài nguyên của chúng ta).
- o Dùng RDFa để thể hiện ngữ nghĩa của dữ liệu thông qua các bộ từ vựng đã được xây dựng.

Hướng phát triển

Dữ liệu chúng ta biểu diễn trên web ngữ nghĩa là dữ liệu đã có ngữ nghĩa, chúng ta có thể liên kết các dữ liệu từ các nguồn dữ liệu từ các tài nguyên ở những nơi khác nhau, để có thể xây dựng hệ thống tổng hợp thông tin ngữ nghĩa từ các nguồn dữ liệu trên web.

Tài Liệu Tham Khảo

Tiếng Việt

- [1] Hoàng Hữu Hạnh (2009). *Web ngữ nghĩa: Những Thách Thức Và Hướng Tiếp Cận Mới*. Tạp chí Khoa Học; Đại Học Huế; Số 48: tr.31-40.
http://hueuni.edu.vn/hueuni/issue_file/48_4.pdf.
- [2] ả guyễn Thị Mỹ Trang, Hoàng Hữu Hạnh (2010). *Xây dựng Ontology cho thư viện số*. Tạp chí Khoa Học. Đại Học Huế; Số 53: tr.127-142.
http://www.hueuni.edu.vn/hueuni/issue_file/53_14.pdf

Tiếng Anh

- [3] Charles McCă (2001), *Semantic Web – History*,
<http://www.w3.org/2001/Talks/0627-semwebh/>.
- [4] Frank Manola, Eric Miller (2004), *RDF Primer*,
<http://www.w3.org/TR/2004/REC-rdf-primer-20040210/> .
- [5] Ben Adida, Mark Birbeck, Ivan Herman (2010), *RDFa Core 1.1*,
<http://www.w3.org/TR/2010/WD-rdfa-core-20100803/>.
- [6] Ben Adida, Mark Birbeck (2008), *RDFa Primer*,
<http://www.w3.org/TR/2008/OTE-xhtml-rdfa-primer-20081014/>.
- [7] Chris Bizer, Richard Cyganiak, Tom Heath (2008), *How to Publish Linked Data on the Web*, <http://sites.wiwiiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/20070727/>.
- [8] Ben Adida, Mark Birbeck, Steven Pemberton (2008), *HTML+RDFa 1.1*,
<http://www.w3.org/TR/2010/WD-rdfa-in-html-20100624/>.