BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

NGUYỄN QUANG MINH

MỘT TIẾP CẬN XÂY DỰNG HỆ THỐNG TỔNG HỢP TIN TỨC THỂ THAO DỰA TRÊN WEB NGỮ NGHĨA

LUẬN ÁN TIẾN SĨ MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG DỮ LIỆU

Hà Nội – 2019

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

NGUYỄN QUANG MINH

MỘT TIẾP CẬN XÂY DỰNG HỆ THỐNG TỔNG HỢP TIN TỨC THỂ THAO DỰA TRÊN WEB NGỮ NGHĨA

NGÀNH: MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG DỮ LIỆU
MÃ SỐ: 9480102

LUẬN ÁN TIẾN SĨ MẠNG MÁY TÍNH
VÀ TRUYỀN THÔNG DỮ LIỆU

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. PGS. TS. NGÔ HỒNG SƠN
2. PGS. TS. CAO TUẤN DŨNG

Hà Nội – 2019

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu khoa học của riêng tôi. Các số liệu, kết quả được công bố với các tác giả khác đều được sự đồng ý của đồng tác giả trước khi đưa vào luận án. Trong quá trình làm luận án, tôi kế thừa thành tựu của các nhà khoa học với sự trân trọng và biết ơn. Các số liệu, kết quả trình bày trong luận án là trung thực và chưa từng được tác giả khác công bố.

Hà Nội, ngày tháng năm 2019

GIẢNG VIÊN HƯỚNG DẪN

PGS. TS Ngô Hồng Sơn

PGS. TS Cao Tuấn Dũng

TÁC GIẢ LUẬN ÁN

Nguyễn Quang Minh

LỜI CẢM ƠN

Tác giả xin bày tỏ lòng biết ơn sâu sắc tới các Thầy hướng dẫn PGS.TS. Ngô Hồng Sơn và PGS.TS. Cao Tuấn Dũng, những người Thầy đã hướng dẫn và giúp đỡ tác giả rất nhiều trong học tập, nghiên cứu khoa học, và thực hiện luận án tiến sĩ. Các Thầy đã luôn khích lệ, động viên và cho tác giả những lời khuyên bổ ích, đặc biệt các Thầy đã chia sẻ thời gian quý báu của mình để giúp tác giả hoàn thành Luận án này.

Bên cạnh đó, tác giả cũng xin gửi lời cảm ơn chân thành tới Ban giám hiệu trường Đại học Bách Khoa Hà Nội, các Thầy/Cô trong Viện Công nghệ thông tin và Truyền thông, các Thầy/Cô ở Bộ môn Truyền thông và mạng máy tính, lãnh đạo và các chuyên viên của Phòng Đào tạo – Bộ phận đào tạo sau đại học đã tạo điều kiện, hỗ trợ và giúp đỡ tác giả trong học tập, trong nghiên cứu và trong công việc suốt thời gian thực hiện Luận án. Sự tận tình của họ khiến tác giả vô cùng xúc động và biết ơn rất nhiều.

Tác giả xin chân thành cảm ơn các Thầy/Cô phản biện, các Thầy/Cô trong Hội đồng các cấp đã trao đổi và cho tác giả nhiều chỉ dẫn quý báu, giúp cho Luận án của tác giả được hoàn thiện, trình bày khoa học và logic hơn.

Tác giả xin chân thành cảm ơn đến nhóm nghiên cứu gồm các bạn: Nguyễn Hoàng Công, Phan Thanh Hiền, Nguyễn Thanh Tâm đã cùng tác giả thực hiện một số nội dung của Luận án.

Tác giả xin bày tỏ lòng biết ơn chân thành tới ban giám đốc Viện Điện tử-Viễn thông đã tạo điều kiện cho tác giả có điều kiện vừa học tập vừa công tác, cảm ơn các đồng nghiệp của bộ môn Điện tử - Kỹ thuật máy tính đã gánh vác một phần công việc giảng dạy trong suốt thời gian tác giả thực hiện Luận án.

Cuối cùng, tác giả xin bày tỏ lòng biết ơn sâu sắc tới toàn thể gia đình, bạn bè, những người thân đã luôn chăm lo, động viên và giúp đỡ tác giả vượt qua mọi khó khăn trong suốt thời gian qua.

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Dạng đầy đủ	Diễn giải
1	CSS	Cascading Style Sheet	Tập tin định kiểu theo tầng
2	FAQ	Frequently Asked Questions	Các câu hỏi thường gặp
3	GATE	General Architecture for Text Engineering	Kiến trúc chung cho kỹ thuật văn bản
4	HTML	Hyper Text Markup Language	Ngôn ngữ đánh dấu siêu văn bản
5	HTTP	Hyper Text Transfer Protocol	Giao thức truyền tải siêu văn bản
6	IRI	Internationalized Resource Identifier	Định danh tài nguyên quốc tế hóa
7	JAPE	Java Annotation Patterns Engine	Công cụ tạo mô hình chú thích Java
8	KBE	Knowledge Base Enrichment	Làm giàu cơ sở tri thức
9	KIM	Knowledge and Information Management	Quản lý tri thức và thông tin
10	NEE	Named Entity Extraction	Trích rút thực thể có tên
11	NER	Named Entity Recognition	Nhận dạng thực thể có tên
12	OKBC	Open Knowledge Base Connectivity	Kết nối cơ sở tri thức mở
13	OWL	Web Ontology Language	Ngôn ngữ ontology trên web
14	QA	Question Answering	Hỏi đáp
15	RDF	Resource Description Framework	Khung mô tả tài nguyên
16	RDFS	RDF Schema	Lược đồ RDF
17	RIF	Rule Interchange Format	Định dạng trao đổi luật
18	SPARQL	SPARQL Protocol and RDF Query Language	Giao thức SPARQL và ngôn ngữ truy vấn RDF
19	TF-IDF	Term Frequency-Inverse Document Frequency	Tần số xuất hiện của 1 từ trong 1 văn bản – Tần số nghịch của 1 từ trong tập văn bản
20	URI	Uniform Resource Identifier	Định danh tài nguyên thống nhất
21	XML	Extensible Markup Language	Ngôn ngữ đánh dấu mở rộng

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN ii
DANH MỤC CÁC TỪ VIẾT TẮT iii
MỤC LỤC iv
DANH MỤC CÁC HÌNH VẼ viii
DANH MỤC CÁC BẢNG ix
MỞ ĐẦU 1
CHƯƠNG 1. KIẾN THỨC NỀN TẢNG VÀ TIẾP CẬN PHÁT TRIỂN HỆ THỐNG TIN TỨC THỂ THAO DỰA TRÊN WEB NGỮ NGHĨA 7
1.1 Giới thiệu về Web ngữ nghĩa 7
1.1.1 Nguồn gốc Web ngữ nghĩa 7
1.1.2 Khái niệm Web ngữ nghĩa 8
1.1.3 Kiến trúc Web ngữ nghĩa 9
1.2 Ontology 10
1.2.1 Định nghĩa 11
1.2.2 Các lĩnh vực ứng dụng và vai trò của ontology 11
1.2.3 Các phương pháp luận phát triển ontology 12
1.2.3.1 Phương pháp luận Methontology 13
1.2.3.2 Phương pháp luận Uschold và King 13
1.2.3.3 Phương pháp luận Grüninger và Fox 14
1.2.4 Các công cụ phát triển ontology 15
1.3 Ngôn ngữ biểu diễn ontology và dữ liệu ngữ nghĩa 15
1.3.1 XML 15
1.3.2 RDF 16
1.3.2.1 Các khái niệm và cú pháp trừu tượng của RDF 16
1.3.2.2 Sử dụng các URI cho các đối tượng thế giới thực 17
1.3.2.3 Phân lớp tường minh các tài nguyên 17
1.3.2.4 Tài nguyên URI, nút trắng, và giá trị hằng 18
1.3.3 RDFS (RDF SCHEMA) 18
1.3.3.1 Các lớp và các thuộc tính 18
1.3.3.2 Miền áp dụng và phạm vi giá trị của các thuộc tính (Domain and Range of Properties) 20
1.3.3.3 Hệ thống kiểu (Type System) 20
1.3.4 OWL (Web Ontology Language) 20
1.3.4.1 Tiên đề và các luật suy diễn kéo theo 21
1.3.4.2 Các tính năng của OWL 21
1.3.4.3 Những tính năng bổ sung trong OWL Full và OWL-DL 22

1.4 Tìm kiếm ngữ nghĩa 22
1.4.1 Các ngôn ngữ truy vấn RDF 22
1.4.2 SPARQL 23
1.4.2.1 Truy vấn SELECT…WHERE 23
1.4.2.2 Truy vấn ASK 23
1.5 Kho dữ liệu ngữ nghĩa mở 24
1.6 Một số lĩnh vực ứng dụng Web ngữ nghĩa 25
1.6.1 Thương mại điện tử 25
1.6.2 Chăm sóc sức khỏe và khoa học đời sống (HCLS) 25
1.6.3 Chính phủ điện tử 25
1.6.4 E-Learning 26
1.7 Một số nghiên cứu Web ngữ nghĩa tiêu biểu 26
1.7.1 Swoogle 26
1.7.2 Dự án ARTEMIS 27
1.7.3 Dartgrid 27
1.7.4 Kho nội dung Web ngữ nghĩa cho nghiên cứu lâm sàng 28
1.7.5 Ứng dụng Web ngữ nghĩa trong lĩnh vực nông nghiệp của tổ chức nông-lương thực Liên hiệp quốc (FAO) 28
1.8 Website và cổng thông tin tin tức có ngữ nghĩa 28
1.8.1 Dự án SWEPT 29
1.8.2 Dự án ARKive 30
1.8.3 Cổng thông tin Esperonto 30
1.8.4 Mondeca ITM 30
1.9 Ứng dụng Web ngữ nghĩa trong lĩnh vực thể thao 30
1.10 Tiếp cận Web ngữ nghĩa xây dựng hệ thống tin tức thể thao 31
1.11 Mô hình kiến trúc hệ thống tổng hợp tin tức thể thao 31
1.11.1 Crawler 32
1.11.2 Ontology thể thao 33
1.11.3 Sinh chú thích ngữ nghĩa 33
1.11.4 Cổng thông tin ngữ nghĩa 34
1.11.5 Mô tơ suy diễn và tìm kiếm ngữ nghĩa 34
1.11.6 Kho dữ liệu ngữ nghĩa 34
1.12 Kết luận chương 35
CHƯƠNG 2. SINH CHÚ THÍCH NGỮ NGHĨA CHO TIN TỨC THỂ THAO 36
2.1 Đặt vấn đề 36
2.2 Chú thích ngữ nghĩa cho tài liệu 37
2.2.1 Khái niệm 37
2.2.2 Các phương pháp tạo chú thích ngữ nghĩa 38
2.2.3 Một số nghiên cứu liên quan 39

2.3 Một phương pháp sinh chú thích ngữ nghĩa cho tin tức thể thao dựa trên ontology và luật trích chọn 40
2.3.1 Tổng quan về phương pháp đề xuất 40
2.3.2 Xây dựng Ontology cho hệ thống 42
2.3.2.1 Ontology PROTON 42
2.3.2.2 Ontology thể thao của hãng BBC 47
2.3.2.3 Xây dựng Ontology BKSport 48
2.3.3 Thu thập và tiền xử lý tin tức 50
2.3.4 Xây dựng cơ sở tri thức thể thao 50
2.3.5 Nhận dạng, trích rút và xác định lớp ngữ nghĩa cho thực thể có tên 51
2.3.5.1 Nhận dạng thực thể có tên trong tin tức như là một thể hiện thuộc cơ sở tri thức 51
2.3.5.2 Phát hiện bí danh của thực thể 52
2.3.5.3 Nhận dạng các thực thể ở mức khái niệm chi tiết 52
2.3.5.4 Cải tiến nhận dạng thực thể có tên ở dạng rút gọn 53
2.3.5.5 Nhận dạng thực thể cùng tên khác kiểu 53
2.3.6 Trích rút "ngữ nghĩa" từ tin tức 53
2.3.6.1 Các ngữ nghĩa bộ ba đơn giản 53
2.3.6.2 Ngữ nghĩa về thực thể quan trọng trong tin tức 53
2.3.6.3 Chú thích ngữ nghĩa về tuyên bố gián tiếp 54
2.3.6.4 Chú thích ngữ nghĩa về tin tức chuyển nhượng 56
2.4 Thực nghiệm 60
2.4.1 Nhận dạng thực thể có tên trong tin tức 61
2.4.2 Trích rút ngữ nghĩa từ tin tức thể thao 65
2.4.3 Đánh giá chung 68
2.5 Kết luận chương 69
CHƯƠNG 3. MỘT PHƯƠNG PHÁP TRUY VẤN TIN TỨC THỂ THAO VỚI NGÔN NGỮ TỰ NHIÊN 70
3.1 Giới thiệu 70
3.2 Các nghiên cứu liên quan 71
3.3 Phân loại câu hỏi đầu vào và cấu trúc truy vấn đầu ra 74
3.3.1 Phân loại câu hỏi 74
3.3.2 Chú thích và truy vấn ngữ nghĩa về tin tức thể thao 75
3.4 Phương pháp chuyển đổi câu hỏi ngôn ngữ tự nhiên sang truy vấn SPARQL 76
3.4.1 Tiền xử lý câu hỏi 77
3.4.2 Phân tích cú pháp 77
3.4.3 Biểu diễn ngữ nghĩa cho câu hỏi 79
3.4.3.1 Mô hình biểu diễn ngữ nghĩa cho câu hỏi 79
3.4.3.2 Chuyển từ cấu trúc ngữ pháp sang biểu diễn ngữ nghĩa 80
3.4.4 Sinh câu truy vấn SPARQL trung gian 84

3.4.4.1 Xác định mệnh đề hỏi 85
3.4.4.2 Xây dựng mệnh đề điều kiện – Mệnh đề WHERE 85
3.4.5 Xác định thực thể, khái niệm và vị từ 87
3.4.5.1 Nhận dạng các lớp 87
3.4.5.2 Nhận dạng thuộc tính 87
3.4.6 Sinh truy vấn SPARQL hoàn chỉnh 88
3.5 Thử nghiệm và đánh giá 89
3.5.1 Kịch bản thử nghiệm và kết quả 89
3.5.2 Nhận xét và đánh giá 91
3.5.2.1 Phân tích cú pháp 91
3.5.2.2 Nhận dạng quan hệ phụ thuộc bộ ba 92
3.5.2.3 Nhận dạng khái niệm và vị từ 92
3.5.2.4 Xử lý nhãn thời gian 92
3.5.2.5 Một số trường hợp đặc biệt chưa xử lý được 92
3.6 Kết luận chương 92
CHƯƠNG 4. GỢI Ý TIN TỨC DỰA TRÊN NGỮ NGHĨA CHO HỆ THỐNG TỔNG HỢP TIN TỨC THỂ THAO 94
4.1 Giới thiệu 94
4.2 Nghiên cứu liên quan 95
4.3 Độ tương đồng giữa các tin 96
4.3.1 Độ tương đồng về ngữ nghĩa 96
4.3.1.1 Quan hệ ngữ nghĩa giữa các thực thể 96
4.3.1.2 Loại thực thể xuất hiện trong tin 100
4.3.1.3 Các chú thích ngữ nghĩa của tin 101
4.3.2 Độ tương đồng về nội dung 102
4.3.3 Thuật toán gợi ý tin tức với độ tương đồng kết hợp 103
4.4 Cài đặt thử nghiệm và đánh giá 104
4.4.1 Kịch bản thử nghiệm 104
4.4.2 Kết quả thử nghiệm và đánh giá 105
4.5 Kết luận chương 106
KẾT LUẬN 107
Các kết quả đạt được của luận án 107
Hướng phát triển 108
DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ CỦA LUẬN ÁN 110
TÀI LIỆU THAM KHẢO 111

DANH MỤC CÁC HÌNH VẼ

Hình 1.1 Kiến trúc Web ngữ nghĩa [59] 9
Hình 1.2 Ví dụ về đồ thị RDF – Tài nguyên được mô tả bằng hình elip, hằng ký tự được mô tả bằng hình chữ nhật. Cạnh có nhãn là URI của vị ngữ sử dụng tiền tố không gian tên 17
Hình 1.3 Ví dụ minh họa một đồ thị RDF nhiều nút 18
Hình 1.4 Định nghĩa FOAF Person như một phần của bảng từ vựng FOAF 19
Hình 1.5 Một phần của Dữ Liệu Liên Kết Mở trên Web, ngày 8 tháng 1 năm 2019 [95] 24
Hình 1.6 Kiến trúc của Swoogle [7] 26
Hình 1.7 Kiến trúc tổng thể của hệ thống BKSport 32
Hình 2.1 Ví dụ về chú thích ngữ nghĩa 38
Hình 2.2 Quá trình chú thích ngữ nghĩa 41
Hình 2.3 Các mô đun của ontology PROTON 43
Hình 2.4 Hệ thống phân lớp của mô đun PROTON System 43
Hình 2.5 Các thuộc tính của mô đun PROTON System 44
Hình 2.6 Tóm lược mô đun ontology PROTON Top 45
Hình 2.7 Tóm lược mô đun ontology PROTON Upper 46
Hình 2.8 Các lớp và thuộc tính của mô đun PROTON KM 47
Hình 2.9 Một phần của ontology thể thao của hãng BBC 47
Hình 2.10 Một phần của ontology BKSport 49
Hình 2.11 Trích rút và xác định lớp ngữ nghĩa cho thực thể có tên 50
Hình 2.12 Một số ánh xạ từ BKSport đến PROTON 51
Hình 2.13 Nhận dạng thực thể có tên trong tin tức thể thao như một thể hiện của cơ sở tri thức 52
Hình 2.14 Các thành phần ngôn ngữ tự nhiên trong mẫu nhận dạng các quan hệ chuyển nhượng 56
Hình 2.15 Các mẫu biểu diễn cụm động từ 57
Hình 2.16 Ví dụ về kết quả nhận dạng đại từ 60
Hình 2.17 Giao diện phần mềm sinh chú thích ngữ nghĩa 62
Hình 2.18 Các thể hiện được nhận dạng bởi KIM và phương pháp đề xuất 63
Hình 2.19 Chú thích ngữ nghĩa được sinh ra với tin tức ở hình 2.18 63
Hình 2.20 Các thể hiện được nhận dạng bởi KIM và phương pháp đề xuất 64
Hình 2.21 Chú thích ngữ nghĩa được sinh ra với tin tức ở hình 2.20 64
Hình 2.22 Chú thích ngữ nghĩa về tuyên bố gián tiếp được trích rút 65
Hình 2.23 Ví dụ về các chú thích nhận dạng đúng 67
Hình 2.24 Ví dụ về các chú thích nhận dạng không đúng 67
Hình 2.25 Ví dụ về các chú thích không được nhận dạng 67
Hình 2.26 Các bộ ba ngữ nghĩa được trích rút là kết quả đầu ra 68
Hình 3.1 Phân loại các câu truy vấn 75
Hình 3.2 Quy trình chuyển đổi câu hỏi từ ngôn ngữ tự nhiên sang SPARQL 77
Hình 3.3 Ví dụ về cây cấu trúc cụm từ trong câu 78
Hình 3.4 Quy trình xác định biến truy vấn 80
Hình 3.5 Xác định các biến thường và ràng buộc quan hệ giữa các biến 81
Hình 3.6 Phương pháp kết hợp hai phụ thuộc theo loại thành một quan hệ bộ ba 82
Hình 3.7 Quy trình xác định ràng buộc về số lượng loại (1) 83
Hình 3.8 Quy trình sinh truy vấn SPARQL trung gian 84
Hình 4.1 Một ví dụ về độ tương đồng giữa hai tin dựa vào các loại thực thể trong tin tức 101
Hình 4.2 Một ví dụ về độ tương đồng giữa hai tin dựa trên các chú thích ngữ nghĩa của tin 102

DANH MỤC CÁC BẢNG

Bảng 2.1. Từ khóa cho các câu tuyên bố gián tiếp 55
Bảng 2.2. Độ chính xác (P) và độ bao phủ (R) của quá trình trích rút từ 150 tin tức thể thao 61
Bảng 2.3. Kết quả trích rút thông tin ngữ nghĩa của thực nghiệm 1 65
Bảng 2.4. Thống kê nhận dạng thực thể có tên và bộ ba của thực nghiệm 2 66
Bảng 2.5. Kết quả bước đầu của thực nghiệm nhận dạng quan hệ ngữ nghĩa 66
Bảng 2.6. Cải thiện hiệu năng của nhận dạng quan hệ ngữ nghĩa 68
Bảng 3.1. Mô hình biểu diễn ngữ nghĩa câu hỏi 79
Bảng 3.2. Một phần của tập các câu hỏi để đánh giá hệ thống đề xuất 91
Bảng 4.1. Độ chính xác gợi ý tin tức trong các trường hợp 105

MỞ ĐẦU

1. Đặt vấn đề

Thế kỉ XXI chúng ta đang sống là một thời đại mà khoa học công nghệ đang ảnh hưởng sâu sắc và thay đổi toàn diện cuộc sống của con người. Đặc biệt khi mà thế giới đã dần chuyển sang nền kinh tế tri thức, việc tiếp cận với những thông tin có giá trị đã trở thành một yếu tố quan trọng quyết định sự thành công của các cá nhân và tổ chức. Bên cạnh đó thông tin còn có mục đích phục vụ nhu cầu mở rộng hiểu biết, đời sống tinh thần của con người, thể hiện rõ nhất ở các tin tức. Tin tức là một loại hình thông tin mà con người đang tiếp cận hàng ngày hàng giờ.

Có nhiều nguồn tin tức từ truyền hình, truyền thanh, báo chí truyền thống và Web. So với các nguồn tin khác, Web có những ưu điểm vượt trội là nhanh, đơn giản, dễ tạo nội dung. Hơn nữa, độc giả hoàn toàn chủ động trong việc lựa chọn thông tin để đọc trên các trang tin điện tử. Vì thế bên cạnh những người dùng Web cá nhân, nhiều hãng tin tức, các công ty truyền thông lớn đã sử dụng Web để phát triển, đưa thông tin cập nhật của họ tới người dùng. Từ đó dẫn đến Web trở thành nguồn tin tức lớn nhất, phong phú, đa dạng và liên tục được cập nhật. Hơn nữa, sự phát triển của các thiết bị công nghệ hiện đại như máy tính xách tay, máy tính bảng, điện thoại thông minh ... đã giúp cho người dùng tiếp cận tin tức trên Web càng dễ dàng, không bị giới hạn về không gian, thời gian. Kết quả là số lượng người dùng tiếp cận thông tin thông qua Web ngày một lớn và tin tức trên Web đã trở thành một xu hướng cho cả người dùng và ngành công nghiệp tin tức hiện đại.

Thể thao nói chung, đặc biệt bóng đá nói riêng, là một lĩnh vực giải trí hấp dẫn, thu hút sự quan tâm của người đọc về các kết quả thi đấu, chuyển nhượng, diễn biến trận đấu, cầu thủ, huấn luyện viên… cũng như các bài bình luận, lời tuyên bố, phát biểu của nhân vật thể thao … trong các giải thi đấu lớn trên thế giới hay châu lục. Hiện nay, những thông tin này đều sẵn có trên Web. Hãng Akamai [1] cho biết lưu lượng internet trung bình của World Cup 2014 là 4.3 Tbps gấp 2.5 lần lưu lượng trung bình của Thế vận hội Mùa đông Sochi 2014 và gấp 7 lần lưu lượng trung bình của World Cup 2010 [2]. Số lượng độc giả truy cập vào các trang Web để đọc tin tức về thể thao gia tăng nhanh chóng. Espn.com đã chào đón khoảng 13 triệu khách khác nhau trong thời gian diễn ra World Cup 2014, tăng 40% so với World Cup 2010 [3]. Vào mỗi đầu mùa bóng mới, Sky Sports thu hút hơn 3 triệu khách khác nhau truy cập các hệ thống tin tức thể thao trên các thiết bị Android và iOS [4]. Yahoo! Sport đón khoảng hơn 1,6 triệu khách khác nhau ghé thăm hàng ngày [5].

Các website thể thao có nội dung phong phú, đa dạng và khổng lồ, nhưng khối lượng thông tin khổng lồ cũng làm người đọc phải mất nhiều thời gian công sức để truy cập những tin tức phù hợp. Họ phải truy cập vào nhiều trang tin khác nhau để tìm, chọn lọc tin tức cũng như thường xuyên phải đọc những tin tức trùng lặp hoặc không cần thiết trong quá trình tìm kiếm của mình. Vì vậy các hệ thống tổng hợp tin tức được xây dựng nhằm giải quyết các khó khăn trên cho người đọc. Với vai trò tổng hợp tin tức từ nhiều nguồn website khác nhau về những lĩnh vực cụ thể nào đó, rồi hiển thị chúng trong một trang Web, các hệ thống ví dụ như Google News hay Baomoi, giúp cho người đọc chỉ với một vài lần truy cập là có thể nhận được đầy đủ thông tin mới nhất về lĩnh vực mình quan tâm thay vì phải truy cập nhiều lần vào các website khác nhau.

Tuy nhiên, khả năng truy cập tin tức trên các website thể thao cũng như các hệ thống tổng hợp tin tức hiện nay vẫn còn một số hạn chế. Các hệ thống này chủ yếu cung cấp chức năng tìm kiếm thông tin theo phương pháp truyền thống dựa trên từ khóa dẫn tới kết quả tìm kiếm không chính xác. Ví dụ, người dùng muốn tìm tin tức với từ khóa "cầu thủ" "chơi hay" "trận kinh điển" có thể nhận kết quả là "cầu thủ Ronaldo đi xem vở kịch kinh điển". Người đọc phải mất nhiều thời gian để xem nội dung các tin tức kết quả trả về bao gồm các tin tức không phù hợp mới có thể tiếp cận được thông tin mình cần. Nguồn gốc của vấn đề nói trên là với mô hình dữ liệu của Web truyền thống, các tin tức hay tài liệu được diễn đạt bởi các thẻ HTML và văn bản

ngôn ngữ tự nhiên. Mô hình này chỉ hướng dẫn máy tính làm thế nào để trình bày thông tin trên một trình duyệt phục vụ cho con người mà không hỗ trợ việc có thể hiểu ý nghĩa của tin tức.

Web ngữ nghĩa [6] là sự mở rộng của Web hiện tại. Ý tưởng của Web ngữ nghĩa là mở rộng các nguyên tắc của Web hiện tại áp dụng trên tài liệu, để chúng hoạt động trên dữ liệu. Công nghệ Web ngữ nghĩa hướng tới phát triển các tiêu chuẩn và công nghệ chung cho phép máy tính hiểu nhiều thông tin trên Web hơn, để chúng có thể hỗ trợ tốt hơn việc khám phá thông tin, tích hợp dữ liệu, và tự động hóa các nhiệm vụ. Ưu điểm của công nghệ Web ngữ nghĩa là cung cấp giải pháp nền tảng để tìm kiếm, trích chọn, tổng hợp thông tin tốt hơn.

Đã có nhiều nghiên cứu cho thấy sự thành công khi ứng dụng công nghệ Web ngữ nghĩa trong giải quyết những bài toán về tìm kiếm thông tin [7] [8] [9], hiển thị thông tin phù hợp ngữ cảnh người dùng [10] và tích hợp dữ liệu [11] [12] [13] [14] trong các lĩnh vực khác nhau như y tế [14] [8], nông nghiệp [12], thương mại điện tử [15], chính phủ điện tử [10] , e-Learning [16] … Tuy nhiên chưa có nhiều nghiên cứu mang tính hệ thống trong việc xây dựng hệ thống tổng hợp tin tức sử dụng tiếp cận này.

Xác định việc nghiên cứu cải thiện, nâng cao chất lượng tìm kiếm, truy cập tin tức là một trong những quan tâm hàng đầu, tác giả lựa chọn hướng nghiên cứu chính là ứng dụng công nghệ Web ngữ nghĩa. Mục tiêu tổng thể là giới thiệu một giải pháp toàn diện hơn cho việc xây dựng các hệ thống tổng hợp tin tức thể thao, đó cũng là lý do luận án này được đặt tên là "Mô hình ngữ nghĩa cho hệ thống tìm kiếm tin tức thể thao".

2. Mục tiêu của luận án

Trên thực tế và cho tới hiện nay, các website tin tức hay các hệ thống tổng hợp tin tức vẫn dựa trên việc sử dụng các hệ quản trị nội dung (CMS) với đặc trưng lưu trữ tin tức sử dụng cơ sở dữ liệu. Các nhà nghiên cứu thường mặc định việc tìm kiếm bằng cách dùng từ khóa, chỉ mục, toàn văn mà vẫn chưa có nhiều nghiên cứu chuyên sâu cho vấn đề tìm kiếm thông tin tốt hơn trong các hệ thống này [17], [18], [19].

Các nghiên cứu về cơ sở lý thuyết và nền tảng công nghệ của Web ngữ nghĩa đã giới thiệu kiến trúc công nghệ của Web ngữ nghĩa còn gọi là Semantic Web Stack, trong đó mỗi tầng liên quan tới một bài toán thành phần cần giải quyết. Cụ thể hơn, mô hình chung được khuyến nghị khi triển khai công nghệ Web ngữ nghĩa cho các hệ thống phần mềm đã được mô tả trong các nghiên cứu [20] [21] [7]. Ở đó các thành phần (hệ thống con) của một hệ thống Web ngữ nghĩa được giới thiệu. Tuy nhiên trong thực tế áp dụng vào các lĩnh vực cụ thể, ngoài ontology là thành phần không thể thiếu và luôn được tập trung xây dựng [22] [14], việc sử dụng các thành phần này được triển khai một cách linh hoạt và có sự khác nhau. Tác giả Ding và các cộng sự trong [7] tập trung vào các thành phần khám phá dữ liệu, tạo chú thích ngữ nghĩa, phân tích dữ liệu và giao diện, trong khi Dogac đề xuất các dịch vụ Web ngữ nghĩa nhằm nâng cao tính liên tác của hệ thống [14]. Thành phần giúp chuyển đổi hay lưu trữ các chú thích ngữ nghĩa là trọng tâm của một số nghiên cứu [13]. Tuy nhiên, chưa có nghiên cứu trình bày về mô hình kiến trúc đầy đủ cho bài toán phát triển hệ thống tin tức thể thao nói chung.

Nghiên cứu về cổng thông tin trang bị công nghệ Web ngữ nghĩa đã có một số kết quả nhất định. Hyvönen [23] đưa ra sự cần thiết của các thành phần metadata, ontology, và các luật trong cổng thông tin. Ahmed và Hmed [24] đã phát triển cổng thông tin ứng dụng Web ngữ nghĩa cho lĩnh vực du lịch. Esperonto và Mondeca ITM [25] [26] là hai nền tảng hỗ trợ xây dựng cổng thông tin ngữ nghĩa có tích hợp một số chức năng như tìm kiếm theo từ khóa, duyệt ontology, quản lý và soạn thảo ontology. Tuy nhiên, chúng còn nhiều hạn chế và gây khó khăn cho việc triển khai trong thực tế như chưa hỗ trợ công cụ suy diễn và giao diện chưa thân thiện.

Các nghiên cứu này chưa đề cập đến vấn đề thu thập, tổng hợp tin tức cũng như các tính năng khai thác thông tin. Các hỗ trợ chủ yếu vẫn là các công cụ để biên tập ontology, hay tạo chú thích ngữ nghĩa, hay thực hiện tìm kiếm một cách thủ công. Vì vậy, một mục tiêu của luận án là đưa ra mô hình kiến trúc cho hệ thống tổng hợp tin tức nói chung và thể thao nói riêng dựa trên nền tảng công nghệ Web ngữ nghĩa. Ở đó làm rõ được vai trò và mối quan hệ giữa các thành phần trong hệ thống và liên hệ tới các bài toán nghiên cứu cụ thể.

Các nghiên cứu nói trên cho thấy để xây dựng một hệ thống ứng dụng công nghệ Web ngữ nghĩa cần giải quyết tốt các bài toán: mô hình hóa ontology, tạo ra các chú thích ngữ nghĩa, thực hiện các tính toán dựa trên suy diễn ngữ nghĩa. Đây cũng là một trong những vấn đề mà luận án quan tâm.

Bài toán về tạo ra các chú thích ngữ nghĩa là tất yếu vì các thế mạnh của Web ngữ nghĩa như tích hợp dữ liệu, tìm kiếm thông tin đều dựa trên một tập các chú thích ngữ nghĩa về các tài nguyên mà hệ thống quan tâm. Các nghiên cứu về sinh chú thích ngữ nghĩa hiện nay đi theo 3 hướng. Hướng thứ nhất là phát triển các công cụ phần mềm để biên tập các chú thích ngữ nghĩa Semantator [27], M-OntoMat Annotizer [28], Annotea [29], Zemanta (http://www.zemanta.com) … Các chú thích ngữ nghĩa được tạo ra một cách thủ công bởi con người có chất lượng tốt nhưng tốn công sức và thời gian. Đối với các hệ thống có dữ liệu khối lượng lớn thường xuyên cập nhật thì phương pháp này gặp khó khăn. Nghiên cứu khác về các phương pháp bán tự động GATE [30], NCBO [31], cTAKE [32] hay tự động như SemTag [33], PANKOW [34] thì tập trung cho lĩnh vực tổng quát hoặc lĩnh vực chuyên biệt khác như sinh học, y tế. Những phương pháp này có một số hạn chế khi triển khai vào lĩnh vực thể thao. Nhiều phương pháp như C-PANKOW [35], KIM [36], AeroDAML [37] mới chỉ tập trung vào việc xác định và gán lớp cho các thực thể có tên, hơn nữa do mục tiêu hướng đến lĩnh vực tổng quát nên các lớp cũng là khái quát như người, địa điểm, thời gian, tiền tệ. Một số phương pháp thì đã trích chọn được quan hệ (thuộc tính) [38] [39] tuy nhiên hiệu quả phụ thuộc vào tri thức của miền ứng dụng. Trong lĩnh vực thể thao để đáp ứng các yêu cầu xử lý thông tin với ngữ nghĩa thì các ngữ nghĩa tạo ra có một số đặc điểm riêng cần được nghiên cứu. Ví dụ, làm thế nào để nhận biết một nhân vật thể thao, biểu diễn các sự kiện hay những kết quả thi đấu … Để đạt được những yêu cầu nói trên cần nghiên cứu phương pháp để nhận dạng được các thực thể có tên trong lĩnh vực thể thao hay sinh ra các chú thích ngữ nghĩa ở dạng bộ ba, bộ bốn.

Một trong những vấn đề điển hình và có ý nghĩa ứng dụng cao của *bài toán tính toán dựa trên suy luận ngữ nghĩa* là tìm kiếm ngữ nghĩa. Trong ngữ cảnh của luận án thì hiệu quả của tìm kiếm ngữ nghĩa đóng vai trò quan trọng trong việc tạo ra giá trị đóng góp về cải thiện độ chính xác của kết quả tìm kiếm của hệ thống tin tức thể thao. Quy trình tìm kiếm ngữ nghĩa gồm 2 bước cơ bản: hình thành câu truy vấn ngữ nghĩa, và thực hiện truy vấn ngữ nghĩa và xử lý kết quả tìm kiếm. Hiện tại bài toán thực hiện truy vấn ngữ nghĩa đã có nhiều kết quả chín muồi, thể hiện ở sự ra đời của các mô tơ tìm kiếm ngữ nghĩa phổ biến trong cộng đồng nghiên cứu như Jena (https://jena.apache.org), Allegrograph (https://allegrograph.com), OpenLink Virtuoso (https://virtuoso.openlinksw.com). Do đó, làm sao tạo ra các truy vấn ngữ nghĩa phù hợp trong lĩnh vực thể thao là một nội dung nghiên cứu cấp thiết.

SPARQL là ngôn ngữ truy vấn ngữ nghĩa được khuyến nghị bởi W3C. Gửi trực tiếp các câu truy vấn SPARQL là hình thức tìm kiếm ngữ nghĩa phổ biến trong các nghiên cứu đầu tiên về vấn đề này [40]. Hiển nhiên là phương pháp này thiếu tính thân thiện người dùng, không phù hợp với những người đọc thông thường. Để hỗ trợ người dùng, [41] tạo ra các giao diện đồ họa dựa trên ontology để hình thành câu tìm kiếm SPARQL. Ngôn ngữ tự nhiên có kiểm soát được sử dụng để tìm kiếm ngữ nghĩa đem lại độ chính xác cao [42] [43], tuy nhiên thiếu sự linh hoạt và chỉ phù hợp cho một miền ứng dụng cụ thể. Tìm kiếm ngữ nghĩa sử dụng ngôn ngữ tự nhiên là một hướng nghiên cứu trong xây dựng các hệ thống hỏi đáp. Từ đó có thể thấy việc tìm ra một hình thức để diễn đạt yêu cầu tìm kiếm thân thiện với người dùng nhưng cho phép tìm kiếm ngữ nghĩa trong hệ thống tổng hợp tin tức là một bài toán nghiên cứu mà luận án có thể đi sâu.

Hệ thống khuyến nghị (Recommender System) là một hệ thống dự đoán sở thích, nhu cầu của người dùng để gợi ý một hoặc nhiều sản phẩm, dịch vụ, thông tin mà người dùng có thể quan tâm. Chính vì vậy trong các hệ thống tin tức, tính năng gợi ý là một tính năng quan trọng. Một trong những tiếp cận phổ biến nhất để xây dựng chức năng này là tiếp cận dựa trên lọc cộng tác. Dựa trên đánh giá của một tập người dùng về các sản phẩm, dịch vụ, cùng với việc so sánh người dùng với tập người dùng nói trên là tư tưởng chính của phương pháp này [44] [45] [46]. Tuy nhiên, các phương pháp dựa trên lọc cộng tác đòi hỏi một số lượng lớn dữ liệu sẵn có về người dùng, điều chỉ có ở các hệ thống lớn đã triển khai trong thực tế. Đó là lý do luận án không đi theo tiếp cận này. Một phương pháp khác, gợi ý dựa theo nội dung, tập trung vào đo

lường đánh giá sự tương đồng giữa nội dung, thuộc tính của các mục cần gợi ý [47] [48]. Trong thời gian gần đây, đã bắt đầu xuất hiện một số nghiên cứu quan tâm đến ngữ nghĩa trong khuyến nghị [49] [50]. Các nghiên cứu này đề xuất độ đo về sự tương đồng ngữ nghĩa giữa các khái niệm xuất hiện trong các văn bản. Đây là một hướng nghiên cứu khá mới và có tiềm năng khai thác khi ứng dụng trong lĩnh vực tin tức.

Mục tiêu nghiên cứu của luận án là xây dựng mô hình, đề xuất phương pháp, kỹ thuật mới… nhằm nâng cao hiệu quả về truy cập tin tức trong hệ thống tổng hợp tin tức. Tiếp cận lựa chọn là dựa trên nền tảng Web ngữ nghĩa. Từ những phân tích về những bài toán cơ bản trong xây dựng hệ thống thông tin dựa trên Web ngữ nghĩa và tình hình nghiên cứu liên quan ở trên, luận án sẽ tập trung giải quyết các mục tiêu nghiên cứu cụ thể như sau:

- Tìm ra một mô hình kiến trúc cho hệ thống tổng hợp tin tức nói chung và thể thao nói riêng dựa trên nền tảng công nghệ Web ngữ nghĩa.
- Nghiên cứu đề xuất các phương pháp sinh ra một cách tự động hoặc bán tự động các siêu dữ liệu còn gọi là chú thích ngữ nghĩa cho các tin tức thể thao. Kết quả của nhiệm vụ này là cơ sở để tiến hành kỹ thuật tìm kiếm ngữ nghĩa trên tin tức. Luận án hướng tới việc sinh ra tự động các chú thích ngữ nghĩa mà nội dung của nó phục vụ cho việc tìm kiếm, đối sánh, giới thiệu, khuyến nghị tin tức. Do đó, các ngữ nghĩa của tin tức thể thao có một số khác biệt (ví dụ, diễn đạt sự kiện xảy ra, con người liên quan, chủ đề liên quan…)
- Thực hiện tìm kiếm ngữ nghĩa trong hệ thống dưới hình thức các câu hỏi bằng ngôn ngữ tự nhiên. Luận án hướng đến giải quyết bài toán chuyển đổi các câu hỏi hay yêu cầu về tin tức dưới dạng ngôn ngữ tự nhiên sang dạng thức truy vấn SPARQL.
- Nghiên cứu phương pháp gợi ý tin tức tới người đọc trên cơ sở sự phù hợp với nội dung của tin tức đang đọc, có khai thác khía cạnh ngữ nghĩa.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của luận án là các bài toán xử lý trên dữ liệu tin tức dựa trên tiếp cận Web ngữ nghĩa. Như vậy luận án sẽ vừa phải tìm hiểu các kiến thức cơ sở lý thuyết nền tảng về Web ngữ nghĩa, vừa phải nắm chắc các phương pháp phân tích xử lý văn bản, cũng như các vấn đề về độ tương quan giữa các văn bản và Ontology.

Luận án được thực hiện trong phạm vi các tin tức tiếng Anh trong lĩnh vực thể thao. Các tin tức ở dạng thức phổ biến nhất là văn bản (text).

Đặt mục tiêu nâng cao hiệu quả của nghiên cứu, luận án xác định không giải quyết bài toán trên phạm vi rộng, bao trùm nhiều lĩnh vực như chính trị, văn hóa, kinh tế … mà chỉ tập trung vào lĩnh vực thể thao. Một nguyên nhân khác là chưa có nhiều nghiên cứu tương tự trong lĩnh vực này. Luận án cũng không xét đến tiếng Việt, lý do là so với tiếng Việt, tiếng Anh có phạm vi áp dụng rộng hơn nhiều.

4. Phương pháp nghiên cứu

Để thực hiện các nội dung nghiên cứu trong luận án, tác giả tiến hành theo phương pháp tiếp cận từ trên xuống, đồng thời kết hợp nghiên cứu lý thuyết với nghiên cứu thực nghiệm.

Về lý thuyết

Bên cạnh nghiên cứu tổng quan các lý thuyết và kỹ thuật cơ bản về Web ngữ nghĩa, các hệ thống tổng hợp tin tức, tác giả phân tích tổng hợp những kết quả nghiên cứu liên quan đã được công bố trong các hội thảo và tạp chí quốc tế. Trên cơ sở đó, tác giả đã xác định được các bài toán nghiên cứu và đề xuất những phương pháp nghiên cứu cho các bài toán nêu trên.

Về thực nghiệm

Luận án tiến hành cài đặt và chạy thực nghiệm, sau đó đánh giá kết quả các phương pháp đã đề xuất trên các tập dữ liệu được xây dựng từ miền lĩnh vực của bài toán.

5. Ý nghĩa khoa học & thực tiễn của luận án, và kết quả nghiên cứu

Ý nghĩa khoa học của các nghiên cứu:

Các phương pháp đề xuất trong luận án cho các bài toán sinh các chú thích ngữ nghĩa cho tin tức thể thao, tìm kiếm với câu hỏi ngôn ngữ tự nhiên, và gợi ý tin tức đều có những đóng góp mới trong phạm vi nghiên cứu tin tức thể thao tiếng Anh:

- Các đề xuất trong nghiên cứu về "sinh chú thích ngữ nghĩa cho tin tức thể thao" có thể làm cơ sở cho các nghiên cứu tiếp theo về vấn đề tạo ra chú thích ngữ nghĩa cho văn bản/tin tức.
- "Phương pháp truy vấn tin tức thể thao với ngôn ngữ tự nhiên" sẽ làm cơ sở cho nghiên cứu chuyển đổi từ câu hỏi ngôn ngữ tự nhiên sang truy vấn ngữ nghĩa sau này.
- Các kết quả trong "Gợi ý tin tức dựa trên ngữ nghĩa" cũng sẽ làm cơ sở cho nghiên cứu các bộ trọng số kết hợp các độ đo về sự liên quan và tương đồng ngữ nghĩa giữa hai văn bản.

Ý nghĩa thực tiễn:

Kết quả nghiên cứu trong luận án có thể được sử dụng trong các hệ thống tổng hợp tin tức thể thao.

Cùng với các kết quả nghiên cứu, luận án cũng đã tiến hành xây dựng mẫu thử nghiệm BKSport và minh họa các thực nghiệm để triển khai ba nghiên cứu trên. Một số nội dung trong kết quả của luận án có thể được áp dụng cho các lĩnh vực khác, đó là những phần không gắn với đặc thù về mặt ngôn từ và diễn tả trong lĩnh vực ứng dụng.

Hệ thống tổng hợp tin tức trong lĩnh vực thể thao mà luận án đã xây dựng được ứng dụng trong thực tế để giúp người dùng tìm kiếm tin tức chính xác hơn và thích hợp với câu hỏi của họ, giúp gợi ý tin tức phù hợp.

Mô hình dựa trên ngữ nghĩa cho hệ thống của luận án tạo điều kiện cho các nghiên cứu về trực quan hóa, tổ chức nội dung của cổng thông tin.

Các kết quả nghiên cứu chính:

- Luận án đề xuất phương pháp để sinh chú thích ngữ nghĩa cho các tin tức thể thao bằng văn bản một cách tự động. Phương pháp này là kết quả của một quá trình của nhiều nghiên cứu với những cải tiến đóng góp vào giải pháp chung, tập trung vào các dạng thức ngữ nghĩa sau:
  - ngữ nghĩa bộ ba đơn giản để diễn tả các sự kiện, các hành động, các chủ đề, các thực thể gắn với tin tức
  - ngữ nghĩa về thực thể quan trọng trong tin tức
  - một số ngữ nghĩa phức tạp như tuyên bố gián tiếp, xử lý đại từ, chuyển nhượng.
- Luận án đề xuất phương pháp chuyển đổi một câu hỏi diễn đạt bằng ngôn ngữ tự nhiên là tiếng Anh thành một truy vấn ngữ nghĩa được biểu diễn dưới dạng thức SPARQL. Truy vấn này là cơ sở để thực hiện tìm kiếm ngữ nghĩa trên hệ thống sử dụng mô tơ tìm kiếm ngữ nghĩa.
- Luận án đề xuất được công thức tính độ tương đồng và liên quan ngữ nghĩa giữa hai tin tức thể thao và sử dụng nó trong phương pháp gợi ý tin tức tới người đọc dựa trên tin tức mà người đó đang đọc.
- Hệ thống tổng hợp tin tức mẫu thử BKSport ứng dụng công nghệ Web ngữ nghĩa đã được triển khai để chứng minh các kết quả nghiên cứu nói trên.

6. Bố cục của luận án

Phần còn lại của luận án được tổ chức thành 4 chương chính. Trong đó, chương 1 giới thiệu kiến thức nền tảng cho các vấn đề được nghiên cứu trong các chương tiếp theo của luận án. Đầu tiên luận án trình bày cơ sở lý thuyết của công nghệ Web ngữ nghĩa phục vụ cho việc sinh chú thích ngữ nghĩa về tin tức và tìm kiếm ngữ nghĩa trong các chương 2, chương 3. Phần tiếp theo của chương tổng hợp thông tin về các nghiên cứu liên quan, đề cập đến các nghiên cứu ứng dụng Web ngữ nghĩa trong nhiều lĩnh vực, và tập trung vào lĩnh vực thể thao. Phần cuối của chương, tác giả khẳng định tiếp cận Web ngữ nghĩa trong xây dựng hệ thống tổng hợp tin tức và đề xuất các nội dung nghiên cứu chính của luận án. Kiến trúc tổng quan của hệ thống tổng hợp tin tức thể thao BKSport dựa trên công nghệ Web ngữ nghĩa cũng được giới thiệu.

Chương 2 trình bày nội dung nghiên cứu về các phương pháp sinh chú thích ngữ nghĩa cho tin tức thể thao dựa trên Ontology, cơ sở tri thức và luật trích chọn. Kết quả thu được là các chú thích ngữ nghĩa với ý nghĩa thể hiện và độ phức tạp khác nhau. Bắt đầu từ phương pháp cơ bản phát hiện kiểu của các thực thể có tên và các bộ ba đơn giản, cho tới chú thích về các tuyên bố gián tiếp và cuối cùng là các ngữ nghĩa phức tạp về chuyển nhượng bóng đá.

Chương 3 đề xuất một phương pháp chuyển đổi câu hỏi ngôn ngữ tự nhiên sang truy vấn SPARQL. Đây là cơ sở để hệ thống tổng hợp tin tức thực hiện tìm kiếm ngữ nghĩa bằng một hình thức tương tác thân thiện với người dùng.

Chương 4 trình bày nội dung kết quả nghiên cứu của phương pháp gợi ý tin tức thể thao có quan tâm đến khía cạnh ngữ nghĩa. Luận án đề xuất độ đo tương đồng giữa hai tin tức trên cơ sở kết hợp độ liên quan ngữ nghĩa và độ tương đồng nội dung.

Cuối cùng là phần kết luận tổng hợp các đóng góp chính của luận án và thảo luận các hướng nghiên cứu trong tương lai.

CHƯƠNG 1. KIẾN THỨC NỀN TẢNG VÀ TIẾP CẬN PHÁT TRIỂN HỆ THỐNG TIN TỨC THỂ THAO DỰA TRÊN WEB NGỮ NGHĨA

Nội dung của chương này trình bày tổng quan về công nghệ Web ngữ nghĩa bao gồm nguồn gốc Web ngữ nghĩa, khái niệm Web ngữ nghĩa, kiến trúc Web ngữ nghĩa, ontology, ngôn ngữ biểu diễn ontology và dữ liệu ngữ nghĩa, tìm kiếm ngữ nghĩa, và kho dữ liệu ngữ nghĩa mở. Các nghiên cứu liên quan trong và ngoài nước về Web ngữ nghĩa cũng được đề cập và phân tích. Đề xuất tiếp cận Web ngữ nghĩa trong xây dựng hệ thống tổng hợp tin tức, các nội dung nghiên cứu chính trong luận án cùng với kiến trúc tổng quan của hệ thống tổng hợp tin tức thể thao BKSport dựa trên công nghệ Web ngữ nghĩa cũng được trình bày cụ thể.

1.1 Giới thiệu về Web ngữ nghĩa

World Wide Web (hay viết tắt là Web) đã trở thành một kho tàng thông tin khổng lồ được tạo ra bởi các tổ chức, cộng đồng và nhiều cá nhân. WorldWideWebSize.com ước tính kích thước của Web trên toàn thế giới cho biết: từ năm 1990 đến năm 2019, Web được lập chỉ mục có chứa ít nhất 5 tỉ trang. Tuy nhiên, do Web ban đầu được thiết kế với mục đích là tạo ra một công cụ giúp con người chia sẻ thông tin một cách dễ dàng, nội dung trên Web hướng tới con người. Vì vậy, Web hiện tại có nhiều hạn chế khi cần được xử lý tự động bởi máy tính. Vấn đề của Web hiện nay đó là người dùng dễ dàng bị lạc, hay phải xử lý một lượng thông tin không hợp lý và không liên quan được trả về từ kết quả tìm kiếm trên Web. Câu hỏi đặt ra là: làm thế nào chúng ta có thể có được kết quả tìm kiếm chính xác một cách nhanh chóng theo những gì mà chúng ta muốn.

Với những hạn chế trên, sự bùng nổ thông tin trên Web đặt ra thách thức mới cho những nhà nghiên cứu. Đó là làm thế nào để khai thác thông tin trên Web một cách hiệu quả. Vấn đề này đã thúc đẩy sự ra đời của ý tưởng Web ngữ nghĩa.

Web ngữ nghĩa không được sinh ra để thay thế toàn bộ Web hiện tại. Mục tiêu của Web ngữ nghĩa là phát triển các tiêu chuẩn và công nghệ chung mà cho phép máy tính hiểu nhiều thông tin trên Web hơn, để chúng có thể hỗ trợ tốt hơn việc khám phá thông tin, tích hợp dữ liệu, và tự động hóa các nhiệm vụ. Thực tế cho thấy rằng Web ngữ nghĩa có thể chứng tỏ những điểm mạnh của mình khi được áp dụng vào những lĩnh vực thông tin bị giới hạn, ví dụ quản lý tri thức, phát triển những dịch vụ Web có ngữ nghĩa.

Với sự hỗ trợ của Web ngữ nghĩa, thông tin mong muốn được tìm ra nhanh hơn và chính xác hơn. Web ngữ nghĩa cũng hỗ trợ tích hợp dữ liệu liên kết từ nhiều nguồn, tìm kiếm động các dữ liệu sẵn có và các nguồn dữ liệu.

1.1.1 Nguồn gốc Web ngữ nghĩa

Tim Berners-Lee là một nhà khoa học máy tính người Anh, nổi tiếng vì phát minh ra World Wide Web với ngôn ngữ đánh dấu siêu văn bản tuy đơn giản nhưng là khuôn dạng đầu tiên cho phép biểu diễn những nội dung giàu thông tin bao gồm văn bản và các dữ liệu đa phương tiện. Ngôn ngữ đánh dấu siêu văn bản là ngôn ngữ đánh dấu mà được các trình duyệt Web sử dụng để trình bày văn bản, hình ảnh, âm thanh, và các tài liệu khác trong các trang web. Tuy nhiên, Tim Berners-Lee thấy nhiều điểm hạn chế của Web hiện tại là nội dung biểu diễn sử dụng HTML mới chỉ hướng đến con người mà chưa thể được hiểu và xử lý tự động bằng máy tính. Từ đó ông đã có ý tưởng thêm ngữ nghĩa vào các trang Web từ gần cuối những năm 1990. Ý tưởng về Web ngữ nghĩa như là phần mở rộng của Web hiện tại trong đó thông tin được xác định rõ ý nghĩa, cho phép máy tính và con người cộng tác với nhau tốt hơn [6].

Nền tảng cho sự ra đời của Web ngữ nghĩa phải nói đến 2 thuật ngữ là RDF và URI. Để gắn siêu dữ liệu phân loại cho các trang Web, nhóm W3C Metadata Activity tạo ra nền tảng PICS (Platform for Internet Content Selection) trong đó các tài nguyên Web được xác định bởi URL

và được cấp các nhãn. URI có khả năng hỗ trợ cho các thực thể trừu tượng, do đó được nhóm Semantic Web Activity đưa ra để thay thế cho các nhãn PICS vốn chỉ đề cập được đến các tài nguyên Web thực (URL).

RDF viết tắt của Resource Description Framework do W3C tạo ra, được sử dụng như một phương pháp chung để mô tả khái niệm hoặc mô hình hóa thông tin về các tài nguyên Web. RDF trở thành mô hình dữ liệu cơ bản cho ontology trên Web, vì với RDF các đối tượng có URI đều có thể được mô tả mà không cần phải có một tài nguyên Web thực sự tồn tại tương ứng.

Từ năm 2001, W3C đã chuẩn hóa những khái niệm cốt lõi của Web ngữ nghĩa cụ thể là RDF, RDFS, OWL (Web Ontology Language), SPARQL, RIF (Rule Interchange Format). Sau 5 năm kể từ ngày phát hành phiên bản SPARQL [51], phiên bản SPARQL 1.1 [52] đã được phát hành vào năm 2013. Phiên bản tiếp theo của OWL [53], ký hiệu là OWL2 [54], đã được công bố vào năm 2012. Phiên bản mới nhất của RIF [55] được công bố vào ngày 5/2/2013.

1.1.2 Khái niệm Web ngữ nghĩa

Năm 2001, Tim Berners-Lee lần đầu tiên giới thiệu chính thức về Web ngữ nghĩa trong một bài báo đăng trên tạp chí Scientific American. Ông đã đưa ra định nghĩa: "*Web ngữ nghĩa là sự mở rộng của Web hiện tại mà ở đó thông tin được định nghĩa một cách rõ ràng, cho phép máy tính và con người có thể hợp tác với nhau tốt hơn*" [6].

Có nhiều nghiên cứu khác nhau với nhiều góc nhìn khác nhau về Web ngữ nghĩa đã được đưa ra bởi các nhà khoa học.

Lassila và các cộng sự [56] mô tả Web ngữ nghĩa như một loạt các tiêu chuẩn, ngôn ngữ mô hình hóa và các sáng kiến phát triển công cụ nhằm chú thích trang Web với siêu dữ liệu được định nghĩa rõ ràng, sao cho các tác nhân thông minh có thể lập luận hiệu quả hơn về các dịch vụ được cung cấp tại các site cụ thể.

Theo Nigel Shadbolt và các cộng sự [57], Web ngữ nghĩa là Web của thông tin hành động – thông tin thu được từ dữ liệu nhờ một lý thuyết ngữ nghĩa để diễn dịch các ký hiệu. Lý thuyết ngữ nghĩa cung cấp một bản kê "ý nghĩa" trong đó các kết nối logic của các thuật ngữ thiết lập khả năng liên tác (interoperability) giữa các hệ thống.

Lee Feigenbaum và các cộng sự [58] phát biểu rằng Web ngữ nghĩa không khác với World Wide Web. Nó là sự nâng cao của Web, cung cấp cho Web tiện ích lớn hơn nhiều. Dựa trên các lược đồ chung, các công cụ Web ngữ nghĩa cho phép liên kết các lược đồ đó, và hiểu các thuật ngữ của chúng để các phần mềm dựa Web ngữ nghĩa của cộng đồng có thể tự động hiểu nhau.

Web ngữ nghĩa là Web của dữ liệu. Ý tưởng của Web ngữ nghĩa là mở rộng các nguyên tắc của Web hiện tại áp dụng trên tài liệu, để chúng hoạt động trên dữ liệu. Khi đó, dữ liệu có thể được truy cập cũng bằng kiến trúc Web chung, ví dụ như là URI. Dữ liệu cũng sẽ được liên kết với nhau giống như những tài liệu Web đã và đang được liên kết. Việc xây dựng Web ngữ nghĩa thành công sẽ tạo ra một khung (framework) cho phép dữ liệu được chia sẻ và tái sử dụng giữa các ứng dụng khác nhau, các doanh nghiệp khác nhau, và cộng đồng khác nhau. Như vậy dữ liệu trong Web ngữ nghĩa sẽ được xử lý tự động/bán tự động cũng như thủ công bằng công cụ.

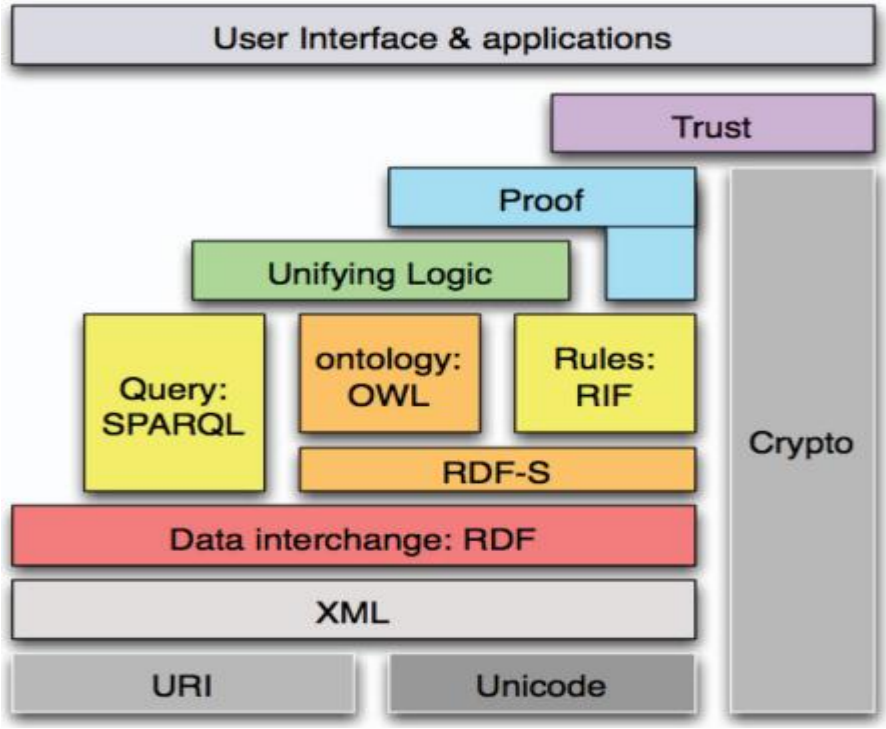
Web ngữ nghĩa có các thành phần quan trọng là ontology, chú thích ngữ nghĩa, và tìm kiếm ngữ nghĩa:

1. Ontology cung cấp vốn từ vựng mô tả các khái niệm và mối quan hệ giữa chúng cho Web ngữ nghĩa. Ontology thể hiện hiểu biết chung về một lĩnh vực mà có thể tái sử dụng và chia sẻ qua các ứng dụng và cộng đồng quan tâm.
2. Chú thích là những giải thích, những ghi chú, sự hiệu đính, sự tham khảo, những bình nghĩa tổng quát hoặc bất kỳ hình thức nào khác của nhận xét bên ngoài được nhúng trong hoặc gắn vào một trang Web hoặc một phần được chọn của tài liệu. Chú thích ngữ nghĩa tổng quát là sự kết hợp của một thực thể dữ liệu với một phần tử của một sơ đồ phân loại, một ontology, hoặc một kho tri thức khác. Chú thích ngữ nghĩa có thể được thực hiện thủ công, tự động hoặc bán tự động.

3. Tìm kiếm ngữ nghĩa là một quy trình tìm kiếm tài liệu dựa trên sự khai thác tri thức lĩnh vực được hình thức hóa bởi một ontology. Nó là một phương pháp cải thiện kết quả tìm kiếm truyền thống bằng cách sử dụng dữ liệu từ Web ngữ nghĩa.

1.1.3 Kiến trúc Web ngữ nghĩa

Hình 1.1 minh họa kiến trúc của Web ngữ nghĩa. Kiến trúc (hoặc ngăn xếp) này đã được đề xuất bởi Berners-Lee và các cộng sự vào năm 2006 [59], và thường được sử dụng để mô tả các thành phần cốt lõi khác nhau của kiến trúc Web ngữ nghĩa. Những thành phần này được khảo sát theo thứ tự từ đáy tới đỉnh của ngăn xếp Web ngữ nghĩa.



Hình 1.1 *Kiến trúc Web ngữ nghĩa [59]*

URI (Uniform Resource Identifier) và IRI (Internationalized Resource Identifier) là một chuỗi ký tự dùng để xác định duy nhất các tài nguyên Web ngữ nghĩa. IRI là tổng quát của URI, IRI chứa các ký tự từ tập hợp ký tự quốc tế (Unicode/ISO 10646) bao gồm Trung Quốc, Nhật Bản, Hàn Quốc v.v. Web ngữ nghĩa cần nhận dạng duy nhất để cho phép thao tác chứng minh được với các tài nguyên ở các lớp trên. URI/IRI là cơ chế cho phép xác định duy nhất các tài nguyên Web ngữ nghĩa. Unicode là cần thiết để có thể biểu diễn các ngôn ngữ toàn cầu. Unicode đảm nhiệm việc biểu diễn và thao tác văn bản trong nhiều ngôn ngữ khác nhau, nó đặc biệt hữu dụng cho trao đổi các ký hiệu.

XML (Extensible Markup Language) là một ngôn ngữ đánh dấu mà cho phép tạo ra các tài liệu về các dữ liệu có cấu trúc. XML là ngôn ngữ định chuẩn công nghiệp trong chuyển giao dữ liệu có cấu trúc trên Web. Tuy nhiên XML mới chỉ hình thức hóa cấu trúc của một tài liệu, nó chưa thể hình thức hóa nội dung của một tài liệu.

RDF (Resource Description Framework) là khung để tạo ra các phát biểu ở dạng các bộ ba <Chủ_thể (S), Đặc_tính (P), Đối_tượng (O)>. Hình thức này cho phép biểu diễn thông tin về các tài nguyên cùng các mối quan hệ của chúng dưới dạng đồ thị. RDF là nền tảng để xử lý siêu dữ liệu, nó đảm bảo tính liên tác giữa các ứng dụng trao đổi thông tin mà máy hiểu được và xử lý được trên Web.

RDFS (RDF Schema) cung cấp một số từ vựng cơ bản để mô hình hóa dữ liệu RDF như lớp và thuộc tính, quan hệ lớp con và thuộc tính con, hạn chế miền và phạm vi... Do đó, RDFS giúp mở rộng ngữ nghĩa cho tài liệu RDF nhờ các cơ chế trên.

OWL (Web Ontology Language) tăng cường RDFS bằng cách đưa ra các tính năng tiên tiến hơn để biểu diễn ngữ nghĩa của các phát biểu RDF. OWL được thiết kế để sử dụng bởi các ứng dụng mà cần xử lý nội dung thông tin thay vì chỉ trình bày thông tin tới người dùng. OWL tạo điều kiện cho máy tính hiểu được nội dung Web hơn rất nhiều so với sự hỗ trợ của XML, RDF, và RDFS. OWL cung cấp các từ vựng bổ sung đi cùng với ngữ nghĩa hình thức để biểu diễn tường minh ý nghĩa của các thuật ngữ trong tập từ vựng và những mối quan hệ giữa chúng. Nó có ba ngôn ngữ con được sắp xếp theo thứ tự tăng dần của khả năng diễn tả như sau: OWL Lite, OWL DL và OWL Full. Các ngôn ngữ ontology trên đều dựa trên cú pháp XML.

SPARQL (SPARQL Protocol and RDF Query Language) là ngôn ngữ để biểu diễn các truy vấn ngữ nghĩa qua nhiều nguồn dữ liệu khác nhau, cho dù dữ liệu được lưu trữ ở định dạng RDF hoặc được xem như RDF nhờ các phần mềm trung gian. Truy vấn dữ liệu ngữ nghĩa trong ontology là một công việc rất quan trọng, đối với các ứng dụng thuần túy khai thác dữ liệu ngữ nghĩa thì không thể thiếu những truy vấn này. Kết quả của truy vấn SPARQL là tập kết quả hoặc đồ thị RDF. Ngôn ngữ truy vấn SPARQL là một ngôn ngữ truy vấn dữ liệu ngữ nghĩa theo chuẩn của hệ thống W3C.

RIF (Rule Interchange Format) là một chuẩn được dùng cho việc trao đổi các luật giữa các hệ thống luật, đặc biệt giữa các mô tơ luật Web. RIF tập trung vào sự trao đổi hơn là cố gắng phát triển một ngôn ngữ luật duy nhất phù hợp cho tất cả. Nguyên nhân ở đây là một ngôn ngữ chuẩn duy nhất không thể đáp ứng được nhu cầu của nhiều mô hình phổ biến khi sử dụng luật trong biểu diễn tri thức và mô hình hóa công việc.

Unifying Logic thực hiện lý luận logic như suy luận sự kiện mới và kiểm tra tính nhất quán.

Proof giải thích rõ các bước lý luận logic của Unifying Logic.

Cryptography bảo vệ dữ liệu RDF thông qua sự mật mã hóa. Nó cũng phê chuẩn nguồn các sự kiện bằng chữ ký số cho dữ liệu RDF.

Trust xác thực độ tin cậy của nguồn tin và các sự kiện được suy ra.

User Interface & applications là giao diện người dùng cho các ứng dụng Web ngữ nghĩa.

1.2 Ontology

Thuật ngữ ontology bắt nguồn từ tiếng Hy Lạp, trong đó onto- (ὄντος) có nghĩa là sự tồn tại và -logy (λογία) có nghĩa là khoa học hay lý thuyết. Như vậy ontology có nghĩa là khoa học về sự tồn tại. Vai trò của ontology là tìm ra thực thể gì đang có trên thế giới, bản chất các thuộc tính của chúng, và chúng có quan hệ với nhau như thế nào. Nhưng nói tóm lại theo cách nhìn của triết học, ontology là *"một môn khoa học về nhận thức, cụ thể hơn là một nhánh của siêu hình học về tự nhiên và bản chất của thế giới, nhằm xem xét các vấn đề về sự tồn tại hay không tồn tại của các sự vật"* [60]. Ontology – bản thể học với ý nghĩa triết học chuyên nghiên cứu về tự nhiên và sự tổ chức, cấu tạo của thế giới thực.

Định nghĩa này bao quát một phạm vi rộng cho phép ontology được hiểu theo nhiều cách. Ví dụ, một ontology có thể là một ngôn ngữ tự nhiên, một mô hình cơ sở dữ liệu cho một bài toán ứng dụng cụ thể hay một hệ thống phân lớp các báo cáo khoa học. Chúng khác nhau ở mức độ diễn tả. Hiển nhiên, việc tìm ra một ontology có khả năng diễn tả cả thế giới hay vũ trụ là không thể.

Hiểu được đặc thù chức năng của ontology trong triết học, khi đứng trước vấn đề cần diễn tả hay mô tả các sự vật hiện tượng thông tin… trong một miền lĩnh vực nào đó, các nhà nghiên cứu trong lĩnh vực CNTT đã vay mượn khái niệm này từ triết học. Mục đích cơ bản của ontology trong CNTT là xây dựng những hệ thống các khái niệm để đặc tả rõ ràng sự nhận thức, hay biểu diễn tri thức của một lĩnh vực cụ thể.

Những giải thích trên khá ngắn gọn và súc tích, tuy nhiên chúng chưa cho phép chúng ta hiểu sâu về ontology. Mục tiếp theo sẽ đi sâu hơn vào từng định nghĩa toàn diện và sâu sắc hơn.

1.2.1 Định nghĩa

Các nhà khoa học đã có nhiều cái nhìn và ý kiến khác nhau về ontology. Họ đã đưa ra nhiều định nghĩa khác nhau về ontology. Sau đây tác giả thống kê lại những định nghĩa đã được thừa nhận rộng rãi như sau:

Neches và các cộng sự [61] định nghĩa ontology như sau: "*Một ontology định nghĩa các thuật ngữ cơ bản và quan hệ bao gồm từ điển của một lĩnh vực nào đó cùng với các luật kết hợp các thuật ngữ với các quan hệ nhằm xác định sự mở rộng cho từ điển*". Định nghĩa này xác định rằng một ontology bao gồm các thuật ngữ cơ bản, các quan hệ giữa các thuật ngữ và các luật để kết hợp các thuật ngữ. Neches cũng cho rằng một ontology bao gồm cả các thuật ngữ được định nghĩa rõ ràng và những tri thức có thể được suy ra từ chúng.

Định nghĩa về ontology được trích dẫn nhiều nhất trong các tài liệu trí tuệ nhân tạo là định nghĩa của Gruber [62]: "*Ontology là một đặc tả rõ ràng cho việc khái niệm hóa trong một lĩnh vực*". Theo tác giả này, thuật ngữ ontology được mượn từ triết học và có nghĩa gốc là sự giải thích có hệ thống về sự tồn tại.

Guarino [63] cho rằng có thể hiểu ontology là một tập hợp các tiền đề logic được thiết kế để giải thích cho ý nghĩa mong đợi của một từ vựng.

Swartout và các cộng sự [64] định nghĩa ontology là một tập thuật ngữ có cấu trúc phân cấp để mô tả một lĩnh vực mà có thể được sử dụng như một nền tảng xương cho một cơ sở tri thức.

Studer và các cộng sự [65] đã định nghĩa ontology là "*Một đặc tả rõ ràng, hình thức của một khái niệm hóa chia sẻ*". Studer và đồng nghiệp cũng giải thích như sau: "*Sự khái niệm hóa có nghĩa là mô hình trừu tượng của các sự vật, hiện tượng trên thế giới được xác định qua các khái niệm liên quan của sự vật, hiện tượng đó. Rõ ràng có nghĩa là các kiểu khái niệm và các ràng buộc giữa chúng là được xác định rõ ràng. Còn hình thức có nghĩa là Ontology phải được hiểu bởi máy tính. Chia sẻ có nghĩa là một ontology không là một thứ riêng tư của một số cá nhân, mà là thứ được sử dụng rộng rãi bởi nhiều người*".

Từ những định nghĩa trên ta có thể đưa ra một khái niệm mang tính chất tổng hợp về ontology như sau. Một ontology là một tập từ vựng bao gồm định nghĩa các khái niệm cơ bản và thuộc tính giữa chúng mà máy tính có thể hiểu được trong một lĩnh vực nào đó. Tập từ vựng này giúp chia sẻ thông tin trong lĩnh vực đó.

1.2.2 Các lĩnh vực ứng dụng và vai trò của ontology

Các cách hiểu khác nhau về ontology cho thấy việc đạt được một sự thống nhất về ngữ nghĩa luôn là vấn đề trong giao tiếp con người. Nghiên cứu và ứng dụng ontology có mục đích cải thiện dần vấn đề trên. Những năm vừa qua, ontology là một chủ đề nghiên cứu được quan tâm trong nhiều lĩnh vực [66], như khoa học đời sống, thiên văn học, toán học, tin học ứng dụng v.v. Đây là những lĩnh vực mà tri thức được thu nhận từ lượng dữ liệu rất lớn được tạo ra. Nhiều công ty và tổ chức nghiên cứu đã ứng dụng ontology và công nghệ Web ngữ nghĩa để quản lý tri thức của họ. Theo Mohammad Mustafa Taye [66], ontology là một chủ đề nghiên cứu phổ biến trong nhiều lĩnh vực như:

1. Web ngữ nghĩa – ontology giúp Web ngữ nghĩa biểu diễn dữ liệu mà máy có thể hiểu được. Nó đóng vai trò quan trọng trong việc trao đổi thông tin giữa các môi trường phân tán.
2. Khám phá dịch vụ Web ngữ nghĩa – ontology đóng vai trò cốt yếu trong việc tìm ra câu trả lời phù hợp nhất cho một truy vấn trong một môi trường kinh doanh điện tử.
3. Trí tuệ nhân tạo – vai trò của ontology ở đây là tạo điều kiện cho việc chia sẻ và tái sử dụng tri thức, cũng như cho phép xử lý qua nhiều chương trình, nhiều dịch vụ, nhiều tác tử, nhiều tổ chức đối với một lĩnh vực cụ thể.
4. Đa tác tử - ontology giữ vai trò quan trọng trong việc cung cấp hiểu biết chung về một tri thức lĩnh vực, do đó nó nâng cao được chất lượng giao tiếp giữa các tác tử.
5. Máy tìm kiếm – ontology đóng vai trò là bộ từ điển thesaurus cho máy tìm kiếm. Nhờ có ontology, máy tìm kiếm có thể trả về thêm các kết quả có chứa các từ đồng nghĩa của một thuật ngữ tìm kiếm. Do đó, chất lượng tìm kiếm được cải thiện.

6. Thương mại điện tử – Giao dịch giữa người bán và người mua được tạo điều kiện dễ dàng hơn nhờ việc sử dụng ontology để mô tả hàng hóa và dịch vụ. Ontology còn giúp giao dịch này được xử lý tự động bởi máy.
7. Khả năng tương tác – ontology cải thiện đáng kể khả năng tương tác giữa các hệ thống ứng dụng phân tán và phi thuần nhất nhờ khả năng tích hợp thông tin vốn có của nó.

Li Ding và các cộng sự [67] cho rằng ứng dụng ontology cho Web ngữ nghĩa đem lại hai lợi ích to lớn sau:

1. Dữ liệu được xuất bản có từ vựng và ngữ pháp chung.
2. Mô tả ngữ nghĩa cho dữ liệu được lưu giữ trong ontology để phục vụ việc suy luận.

Tác giả này cũng cho rằng ontology có ba ứng dụng đối với Web ngữ nghĩa như sau:

1. Khám phá dịch vụ ngữ nghĩa – ontology được sử dụng để mô tả các dịch vụ dữ liệu khác nhau trong mạng ad-hoc, để lý luận về khả năng của thiết bị cảm biến v.v. Một ứng dụng nổi bật đó là ontology Service cùng với các tính năng mở rộng của nó.
2. Tích hợp hồ sơ cá nhân dựa trên ontology – ontology được sử dụng để xây dựng một CSDL quy mô mạng toàn cầu về hồ sơ cá nhân. Một ứng dụng nổi bật đó là ontology FOAF được đánh giá là có tầm nhìn xa.
3. Suy diễn dựa trên logic mô tả cho các cảm biến thích nghi – ontology được sử dụng để suy luận các trạng thái của thiết bị cảm biến dựa trên các tiền đề có trong OWL-DL. Một ứng dụng nổi bật đó là ontology Sensor State được đánh giá cao về khả năng suy luận.

Theo Ian Horrocks [68], ontology được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau như sinh học, y học, địa lý học, địa chất học, nông nghiệp và quốc phòng. Lĩnh vực khoa học đời sống cho thấy những thành tựu to lớn của việc ứng dụng của ontology với các sản phẩm nổi bật trong lĩnh vực y sinh học như SNOMED, GO and BioPAX, Foundational Model of Anatomy (Mô Hình Nền Tảng Của Giải Phẫu Học), và the National Cancer Institute thesaurus (bộ từ điển thesaurus Viện Ung Thư Quốc Gia). Ontology cũng được sử dụng rộng rãi để tạo điều kiện thuận lợi cho việc chia sẻ và tích hợp thông tin. Trong các ứng dụng tích hợp thông tin, ontology được dùng để:

1. cung cấp vốn từ vựng được định nghĩa hình thức và có thể mở rộng để sử dụng trong các chú thích ngữ nghĩa,
2. mô tả cấu trúc các nguồn hiện có và thông tin chúng lưu trữ, và
3. cung cấp một mô hình chi tiết về lĩnh vực để đáp ứng được những truy vấn nâng cao. Những truy vấn như vậy có thể được trả lời bằng cách sử dụng chú thích ngữ nghĩa và tri thức có cấu trúc để truy tìm và kết hợp thông tin từ nhiều nguồn khác nhau.

Nhóm tác giả Aarti Singh và Poonam Anand [69] đưa ra những lý do sau đây của việc phát triển ontology:

- Để chia sẻ sự hiểu biết chung về cấu trúc của thông tin giữa con người hoặc các tác tử phần mềm.
- Để cho phép tái sử dụng các tri thức của một lĩnh vực cụ thể
- Để làm cho các giả định trong một lĩnh vực được tường minh
- Để tách tri thức lĩnh vực khỏi tri thức thao tác
- Để phân tích tri thức lĩnh vực

1.2.3 Các phương pháp luận phát triển ontology

Trong những năm gần đây, có nhiều phương pháp luận khác nhau được đưa ra để hỗ trợ việc phát triển ontology. Những phương pháp luận cổ điển bao gồm Cyc [70], Uschold và King [71], Grüninger và Fox [72], KACTUS [73], và Methontology [74]. Những phương pháp luận này cung cấp những hướng dẫn có cấu trúc và chung nhất giúp đẩy nhanh quá trình phát triển và cải thiện được chất lượng của các ontology kết quả. Trong bài báo "Apport de l'ingénierie ontologique aux environnements de formation à distance" [75], Psyché và cộng sự đã phân loại các phương pháp luận xây dựng ontology cổ điển thành năm nhóm:

- xây dựng từ đầu, ví dụ Uschold và King [71]

- tích hợp với các ontology khác, ví dụ Gruninger và Fox [72]
- tái kỹ nghệ
- xây dựng cộng tác
- đánh giá ontology

1.2.3.1 Phương pháp luận Methontology

Phương pháp luận thiết kế ontology phổ biến nhất là "Methontology". Khung Methontology cho phép xây dựng các ontology ở mức tri thức và giới thiệu về: quy trình phát triển ontology, vòng đời ontology, và các kỹ thuật cụ thể để thực hiện mỗi hoạt động.

Methontology bao gồm các hoạt động sau đây để phát triển một ontology:

1. Đặc tả: nhiệm vụ thứ nhất của pha này là xác định mục đích của ontology, bao gồm người dùng mà nó hướng đến, các kịch bản sử dụng của nó, mức độ hình thức yêu cầu… Nhiệm vụ thứ hai là xác định phạm vi của ontology. Cụ thể hơn cần xác định tập thuật ngữ được ontology biểu diễn, đặc tính và độ chi tiết được yêu cầu của tập thuật ngữ này. Kết quả của pha này là một tài liệu đặc tả ontology ở dạng ngôn ngữ tự nhiên.
2. Thu nhận tri thức: giai đoạn này phần lớn được thực hiện song song với giai đoạn đặc tả (1). Vai trò của các cuộc phỏng vấn với chuyên gia và phân tích văn bản được quan tâm đặc biệt. Tuy nhiên, hoạt động này không tuân theo một quy tắc cứng nhắc ví dụ như là phải khai thác từ một loại nguồn tri thức và sử dụng phương pháp khơi gợi nào đó.
3. Khái niệm hóa: các thuật ngữ miền lĩnh vực được định nghĩa là các khái niệm, các thể hiện, các quan hệ ở dạng động từ hoặc các thuộc tính, và mỗi thuật ngữ đó được biểu diễn bằng một dạng biểu diễn phi hình thức khả dụng.
4. Tích hợp: nhằm đạt được một số đồng nhất trên các ontology và trên các định nghĩa từ các ontology khác. Hoạt động này giúp đẩy nhanh tiến độ xây dựng các ontology vì giúp tái sử dụng các định nghĩa từ các ontology khác.
5. Phát triển: ontology được biểu diễn hình thức bằng một ngôn ngữ nào đó, chẳng hạn như RDF hay OWL.
6. Đánh giá: Methontology chú trọng vào giai đoạn này. Hoạt động này sử dụng các kỹ thuật dùng trong thẩm định và kiểm chứng các hệ cơ sở tri thức, ví dụ như tìm kiếm sự không đầy đủ, thiếu nhất quán và dư thừa trong ontology …
7. Tư liệu hóa: là đối chiếu các tài liệu có được từ các hoạt động khác.

Vòng đời của một ontology

Các hoạt động trên được sắp xếp thứ tự trong một chu trình được gọi là vòng đời của một ontology. Một ontology đi qua các trạng thái sau: đặc tả, khái niệm hóa, hình thức hóa, tích hợp, phát triển. Cuối cùng, các ontology vào trạng thái bảo trì. Pha thu nhận tri thức, đánh giá và tài liệu hóa được thực hiện trong toàn bộ vòng đời.

Giống như Tove, khía cạnh đặc biệt nhất của Methontology là tập trung vào bảo trì. Sự khác biệt chính giữa hai phương pháp luận này là ở chỗ Methontology tập trung vào việc giải quyết toàn diện giai đoạn bảo trì của chu trình sống của ontology, trong khi Tove sử dụng các kỹ thuật hình thức hơn để giải quyết một số lượng hạn chế hơn về vấn đề bảo trì.

1.2.3.2 Phương pháp luận Uschold và King

Dựa trên kinh nghiệm xây dựng Enterprise ontology – một ontology cần thiết trong các quy trình mô hình hóa doanh nghiệp, tại Đại học Edinburgh các tác giả Uschold và King đã đưa ra một phương pháp luận để phát triển ontology. Phương pháp luận này gồm các giai đoạn như sau:

1. Xác định mục đích: nhiệm vụ của giai đoạn này là làm rõ lý do ontology cần được xây dựng và những ứng dụng mục tiêu ontology hướng tới là gì. Ngoài ra, người ta cũng xác định mức độ hình thức mà ontology cần mô tả.

2. Xác định phạm vi: bằng cách dùng các kịch bản và câu hỏi kiểm chứng khả năng ở dạng phi hình thức, giai đoạn này đưa ra một đặc tả yêu cầu và phác thảo đầy đủ phạm vi thông tin mà ontology mô tả.
3. Xây dựng ontology: giai đoạn này xác định các khái niệm và các mối quan hệ chính trong miền lĩnh vực quan tâm, tạo ra các định nghĩa văn bản rõ ràng chính xác cho các khái niệm và các mối quan hệ, xác định các thuật ngữ để chỉ các khái niệm và các mối quan hệ. Sau đó là xem xét khả năng tích hợp với các ontology có sẵn.
4. Hình thức hóa bằng cách tạo ra "mã", các định nghĩa hình thức và các tiên đề của các thuật ngữ trong đặc tả. Công việc này bao gồm việc biểu diễn tường minh những tri thức thu được trong một ngôn ngữ hình thức nào đó.
5. Đánh giá hình thức: việc đánh giá trong giai đoạn này có thể sử dụng các tiêu chuẩn đánh giá cụ thể cho một ontology riêng biệt, hoặc sử dụng các tiêu chuẩn đánh giá khái quát [76] cho đa số các ontology.

Nói chung, với hầu hết các phương pháp luận phát triển hệ cơ sở tri thức gần đây, phương pháp tiếp cận Uschold & King phân biệt giữa pha phi hình thức và hình thức trong việc xây dựng ontology. Pha phi hình thức liên quan đến việc xác định khái niệm quan trọng sau đó đưa ra định nghĩa văn bản cho các khái niệm và các mối quan hệ, và sử dụng các kỹ thuật thu nhận tri thức sẵn có.

Nhược điểm của phương pháp luận Uschold & King là không đủ chi tiết để mô tả chính xác các kỹ thuật sử dụng và các thao tác.

1.2.3.3 Phương pháp luận Grüninger và Fox

Từ kinh nghiệm xây dựng các ontology trong lĩnh vực mô hình hóa các hoạt động và các quy trình nghiệp vụ, Grüninger và Fox đã đề xuất ra phương pháp luận Tove (Toronto Virtual Enterprise) [72] [77] trong dự án cùng tên. Các ontology này bao gồm: Enterprise Design Ontology, Project Ontology, Scheduling Ontology, và Service Ontology. Dưới đây là các giai đoạn chính:

1. Biên tập các kịch bản: đây là điểm bắt đầu của việc xây dựng ontology. Các kịch bản thường là những vấn đề gặp phải trong một tổ chức mà không được giải quyết thỏa đáng bởi các ontology sẵn có đi kèm với các giải pháp mang tính trực giác tương ứng. Nó thường ở dạng các câu chuyện kể lại hoặc các ví dụ.
2. Đặt các câu hỏi kiểm chứng khả năng ở dạng phi hình thức: dựa trên kịch bản ở giai đoạn (1), các yêu cầu đối với ontology được mô tả ở dạng những câu hỏi phi hình thức (tới lúc này chúng vẫn còn chưa được thể hiện bằng ngôn ngữ hình thức của ontology). Một ontology phải có khả năng biểu diễn những câu hỏi này bằng hệ thống thuật ngữ của nó, và có thể mô tả câu trả lời cho những câu hỏi này bằng tiền đề và định nghĩa của nó.
3. Đặc tả thuật ngữ: từ các câu hỏi kiểm chứng phi hình thức, các thuật ngữ của ontology như các khái niệm, thuộc tính và mối quan hệ được diễn tả sử dụng một hệ hình thức nào đó.
4. Hình thức hóa các câu hỏi kiểm chứng khả năng: giai đoạn này được thực hiện đơn giản bởi việc dùng các thuật ngữ hình thức của ontology trong biểu diễn câu hỏi kiểm chứng khả năng phi hình thức.
5. Đặc tả tiên đề: các tiên đề đặc tả định nghĩa và các ràng buộc về mặt diễn dịch của các thuật ngữ được đưa ra ở dạng logic bậc nhất. Các tiên đề này là điều kiện cần và đủ để diễn đạt các câu hỏi kiểm chứng khả năng và các đáp án tương ứng.
6. Thiết lập các điều kiện về tính đầy đủ của ontology: giai đoạn này định nghĩa các điều kiện mà theo đó các đáp án nói trên là đầy đủ. Vì vậy các điều kiện này được gọi là định lý về tính đầy đủ.

Điểm nổi bật của phương pháp luận Tove là chú trọng vào việc đánh giá ontology sử dụng các định lý trên. Những định lý này rất hữu ích trong một số nhiệm vụ bảo trì ontology, ví dụ đánh giá khả năng mở rộng của một ontology.

1.2.4 Các công cụ phát triển ontology

Ontology được xây dựng nhằm mục đích nắm bắt tri thức một cách hình thức và theo cách chung nhất. Nó có thể được tái sử dụng và chia sẻ qua các ứng dụng và các nhóm người. Ontology đóng một vai trò quan trọng trong Web ngữ nghĩa, trích chọn thông tin, trí tuệ nhân tạo, xử lý ngôn ngữ tự nhiên, quản lý tri thức, vv… Xây dựng ontology là một nhiệm vụ đầy thử thách. Phương pháp phổ biến là xây dựng thủ công ontology rất tốn thời gian và phức tạp. Có rất nhiều công cụ có sẵn để xây dựng ontology. Những công cụ này hỗ trợ quá trình phát triển ontology, giúp người dùng xây dựng nên các ontology và cần phải chọn ra công cụ thích hợp cho mục đích này. Mục này khảo sát và phân tích so sánh các công cụ sẵn có cho việc xây dựng ontology.

Có hai loại công cụ xây dựng ontology chủ yếu:

1. Các công cụ soạn thảo ontology: cho phép người dùng định nghĩa các khái niệm mới, các mối quan hệ mới, và các thể hiện mới. Các công cụ này thường bao gồm các trình duyệt đồ họa, chức năng tìm kiếm, bộ kiểm tra ràng buộc. Một số ví dụ điển hình của những công cụ này là Protégé [78], OntoEdit [79], WebODE [80].
2. Các công cụ ánh xạ, căn chỉnh và trộn ontology: đây là những công cụ giúp người dùng tìm thấy những điểm tương tự và những điểm khác biệt giữa các ontology nguồn. Chúng hoặc xác định một cách tự động sự tương ứng tiềm năng hoặc cung cấp môi trường cho người sử dụng tìm và xác định các tương ứng này, hoặc cả hai. Những công cụ ánh xạ này thường là phần mở rộng của các công cụ phát triển. Một số ví dụ điển hình của những công cụ này là PROMPT, ONION, Chimaera [81] [82].

1.3 Ngôn ngữ biểu diễn ontology và dữ liệu ngữ nghĩa

1.3.1 XML

XML được phát triển bởi XML Working Group (ban đầu là Ban Biên Tập Đánh Giá SGML). Nhóm này được thành lập dưới sự bảo trợ của W3C vào năm 1996. XML, là chữ viết tắt của Extensible Markup Language, đã trở thành khuyến nghị W3C vào ngày 10/02/1998. XML không phải là một sự thay thế cho HTML, nó là sự bổ sung thông tin cho HTML. XML được thiết kế để cấu trúc hóa, trao đổi, chia sẻ, vận chuyển và lưu trữ dữ liệu, tập trung vào dữ liệu là gì. Trong khi đó, HTML được thiết kế để hiển thị dữ liệu, tập trung vào dữ liệu trông như thế nào. Ngôn ngữ XML không có các thẻ được tiền định nghĩa, các thẻ XML cũng như cấu trúc tài liệu XML được định nghĩa bởi tác giả của tài liệu XML đó. Khi ta cần hiển thị dữ liệu động trong tài liệu HTML, sẽ mất rất nhiều công sức để chỉnh sửa tài liệu HTML này mỗi khi dữ liệu thay đổi. Với XML, dữ liệu được lưu trữ trong các tập tin XML riêng biệt. Do đó, những thay đổi trong dữ liệu nằm dưới sẽ không ảnh hưởng tới việc hiển thị và bố trí với HTML/CSS. Dữ liệu XML là độc lập với phần cứng và phần mềm, do đó nó dễ dàng được chia sẻ và tái sử dụng bởi các ứng dụng khác nhau. Việc trao đổi dữ liệu giữa các hệ thống không tương thích trên internet được giảm đáng kể về độ phức tạp cũng như về chi phí thời gian khi sử dụng dữ liệu XML.

Lược đồ XML giúp cấu trúc một tài liệu XML. Cấu trúc này được xác định bằng một danh sách các phần tử hợp lệ. Trong XML, các tên của phần tử được định nghĩa bởi các nhà phát triển. Điều này có thể dẫn đến xung đột khi kết hợp các tài liệu XML từ nhiều ứng dụng XML khác nhau. Không gian tên XML giúp giải quyết những xung đột về tên này bằng cách sử dụng các tiền tố tên duy nhất. XML có luật cú pháp đơn giản mạnh mẽ giúp tạo nên các tài liệu XML có dạng cấu trúc cây, nhưng không áp đặt các ràng buộc ngữ nghĩa lên ý nghĩa của các tài liệu này. Nhiều ngôn ngữ mới dựa trên Internet đã được tạo ra với XML như WSDL (mô tả các dịch vụ Web có sẵn), WAP và WML (ngôn ngữ đánh dấu cho các thiết bị cầm tay), RSS (ngôn ngữ cho nguồn cấp dữ liệu tin tức), RDF và OWL (mô tả các tài nguyên và ontology), SMIL (mô tả đa phương tiện cho web), XHTML (một phiên bản chặt chẽ hơn, đầy đủ hơn và chính xác hơn dựa trên XML của HTML) v.v.

Trong khi XML là hoàn toàn phù hợp cho việc trao đổi dữ liệu có cấu trúc, có ba khía cạnh quan trọng mà nó còn thiếu. Thứ nhất, các phần tử lược đồ, các thuộc tính, và các thực thể được định nghĩa không bổ sung thêm ngữ nghĩa cho tên của chúng. Ví dụ, một thuộc tính có tên là tempValue có thể có nghĩa là một giá trị nhiệt độ hoặc biểu thị một giá trị tạm thời. Để diễn dịch dữ liệu XML một cách chính xác, bên cạnh suy diễn của con người thì người ta còn thường cần đến một số tư liệu bổ sung cho lược đồ XML. Thứ hai, XML có khả năng hạn chế khi mô tả các mối quan hệ giữa các phần tử có liên quan tới các đối tượng. Mặc dù nó có thể sử dụng các thuộc tính ID và IDREF để định danh các phần tử và tham chiếu tới các phần tử khác, nhưng những sự tham chiếu này không có bất kỳ ý nghĩa kết hợp đặc biệt nào. Thứ ba, XML dựa trên giả định thế giới đóng, và do đó nó không thể thêm thông tin bổ sung cho các tài liệu XML đã tồn tại, và hơn nữa nó không thể kết hợp với các tập thông tin XML phân tán.

Do đó, những ngôn ngữ đánh dấu Web mạnh mẽ hơn XML là cần thiết để thực hiện các nhiệm vụ xử lý thông tin phức tạp hơn. Một cách để giải quyết vấn đề này là liên kết ý nghĩa máy có thể xử lý được với các thẻ sử dụng các kỹ thuật biểu diễn tri thức như là RDF, RDFS hay OWL.

1.3.2 RDF

RDF (Resource Description Framework) là mô hình dữ liệu cốt lõi của tất cả các ứng dụng dựa trên Web ngữ nghĩa. Các đặc tả RDF hiện nay được chia thành sáu chuẩn khuyến nghị được đề xuất bởi W3C: nhập môn RDF (RDF Primer), các khái niệm và cú pháp trừu tượng của RDF (RDF Concepts and Abstract Syntax), đặc tả cú pháp RDF/XML (RDF/XML Syntax Specification), ngữ nghĩa RDF (RDF Semantics), lược đồ RDF (RDF Schema), và các ca kiểm thử RDF (RDF Test Cases). Các mục dưới đây trình bày một số nội dung quan trọng của RDF. Đó là các khái niệm cơ bản, làm thế nào để sử dụng RDF hiệu quả, cách thức để định nghĩa các từ vựng sử dụng RDF Schema, và các ứng dụng sử dụng RDF.

1.3.2.1 Các khái niệm và cú pháp trừu tượng của RDF

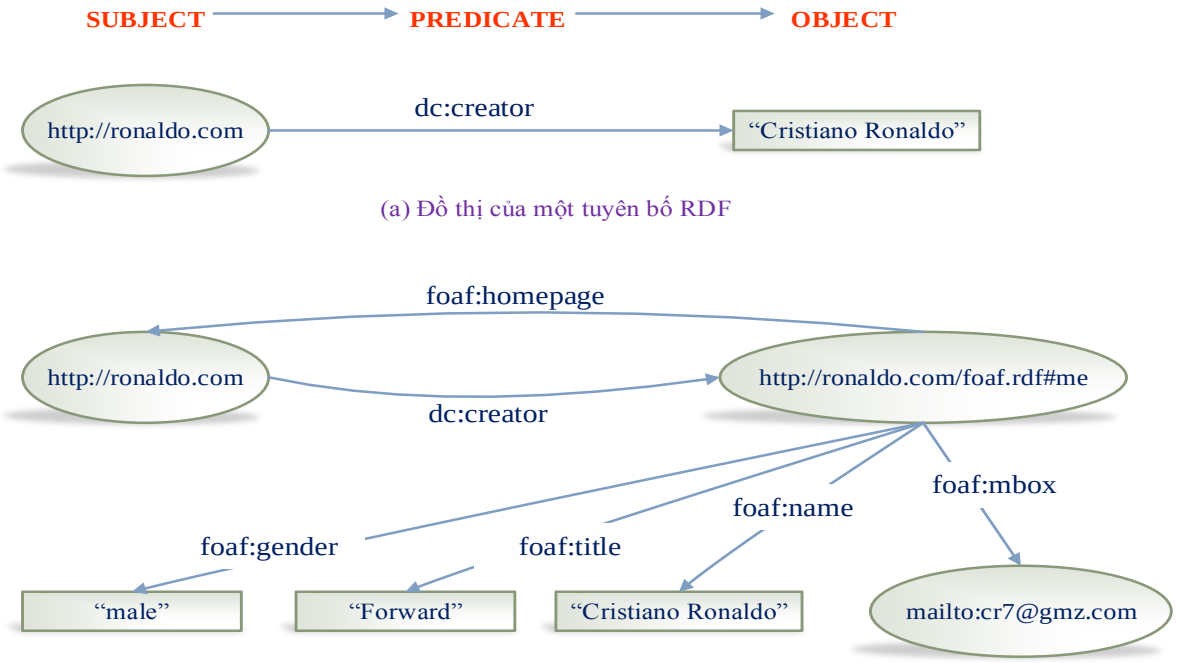
Sức mạnh của RDF rõ ràng là nằm ở mô hình dữ liệu cốt lõi đơn giản dựa trên tập các tuyên bố RDF có dạng (S, P, O), với S biểu thị chủ ngữ, P biểu thị vị ngữ, và O biểu thị tân ngữ tương tự như câu ngôn ngữ tự nhiên. Một tuyên bố như:

Trang web tại http://ronaldo.com/ đã được tạo ra bởi Cristiano Ronaldo.

có thể được diễn đạt trong đồ thị RDF thể hiện trong hình 1.2a. Trong ví dụ này, chủ ngữ là tài nguyên với URI http://ronaldo.com/, vị ngữ là dc:creator (một thuộc tính Dublin Core được tiền định nghĩa với URI http://purl.org/dc/elements/1.1/creator), và hằng ký tự "Cristiano Ronaldo" là tân ngữ.

Hình 1.2b mô tả đồ thị tương ứng với sáu tuyên bố sau đây (được thể hiện bằng định dạng Turtle):

```
@prefix foaf:   <http://xmlns.com/foaf/0.1/> .
@prefix ex:     <http: //example.org#> .
<http://ronaldo.com/>                   dc:creator      <http://ronaldo.com/foaf.rdf#me> .
<http://ronaldo.com/foaf.rdf#me>        foaf:homepage   <http://ronaldo.com/> .
<http://ronaldo.com/foaf.rdf#me>        foaf:name       "Cristiano Ronaldo" .
<http://ronaldo.com/foaf.rdf#me>        foaf:mbox       <mailto: cr7@gmz.com> .
<http://ronaldo.com/foaf.rdf#me>        foaf:title      “Forward” .
<http://ronaldo.com/foaf.rdf#me>        foaf:gender     “male” .
```



Hình 1.2 *Ví dụ về đồ thị RDF – Tài nguyên được mô tả bằng hình elip, hằng ký tự được mô tả bằng hình chữ nhật. Cạnh có nhãn là URI của vị ngữ sử dụng tiền tố không gian tên*

Mỗi một tuyên bố được biểu diễn trong đồ thị bằng một cung có hướng. Hai ký hiệu foaf: và ex: là các tiền tố không gian tên được tạo ra theo cú pháp tổng quát sau:

@prefix [prefix-name]: <[namespace-uri]>

Do đó, foaf:name là dạng rút gọn của URI http://xmlns.com/foaf/0.1/name. FOAF là ontology Friend-of-a-Friend, cung cấp bộ từ vựng để mô tả người và các mạng xã hội giữa người với người. Mỗi không gian tên xác định duy nhất một từ vựng RDF cụ thể.

Để thêm các thông tin về thực thể Cristiano Ronaldo, cần có các tài nguyên khác bổ sung ngữ nghĩa cho chuỗi ký tự "Cristiano Ronaldo". Chỉ có tài nguyên thì mới được dùng làm chủ ngữ trong các câu tuyên bố RDF.

URI http://ronaldo.com/foaf.rdf#me giới thiệu một tài nguyên RDF mới, đại diện cho nhân vật thể thao có foaf:name dẫn đến "Cristiano Ronaldo" và có URI foaf:mbox dẫn đến <mailto: cr7@gmz.com>.

1.3.2.2 Sử dụng các URI cho các đối tượng thế giới thực

Nguyên lý cơ bản là tất cả mọi thứ có thể được mô tả bởi người nào đó trên Web sẽ nhận được một URI và để có thể lấy thông tin về nguồn tài nguyên, URI của nó là phân giải được bởi các client HTTP [83]. Trong một số trường hợp, có thể xảy ra khả năng là một tài nguyên không đòi hỏi phải có một URI tường minh. Để giải quyết vấn đề này, RDF hỗ trợ khái niệm nút trắng, đó là nút tài nguyên mà không có URI toàn cục duy nhất.

1.3.2.3 Phân lớp tường minh các tài nguyên

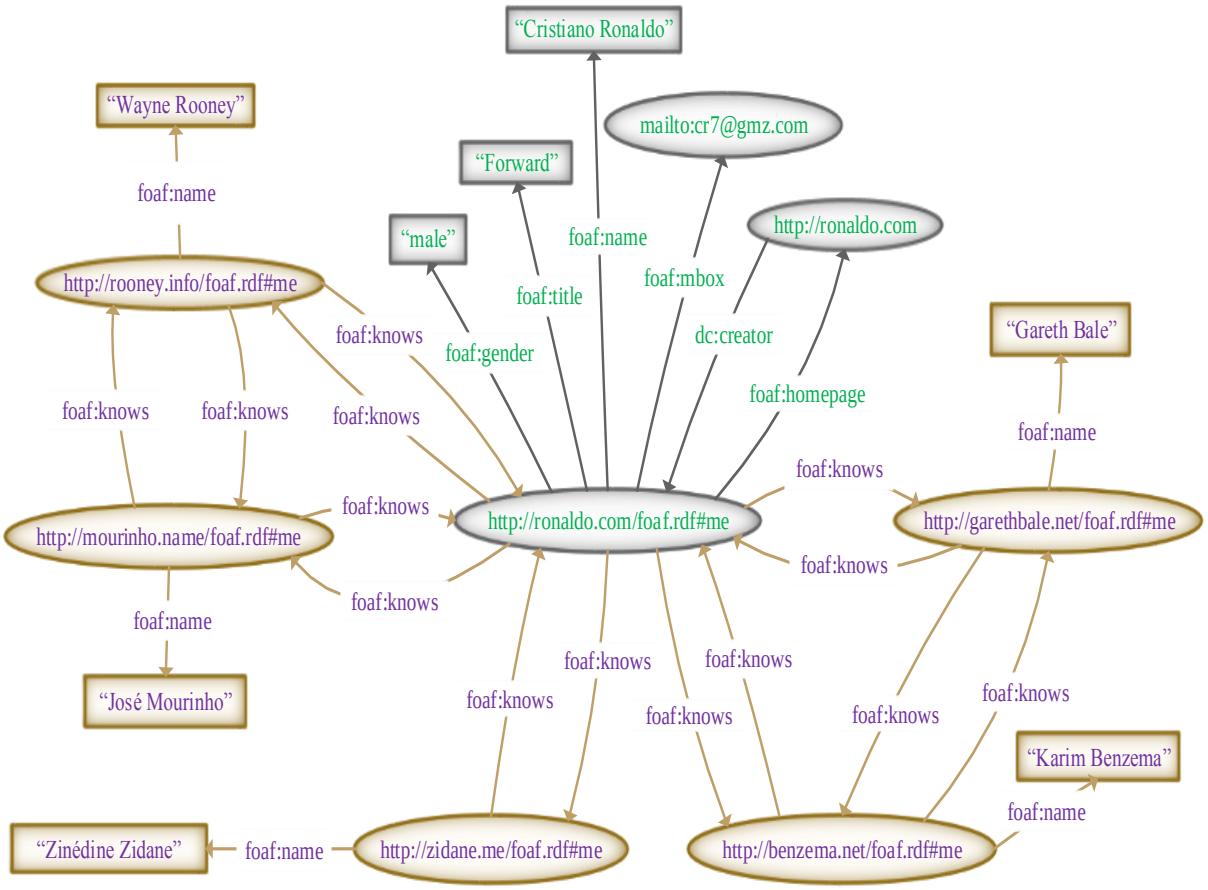
Để diễn tả một tài nguyên nào đó là thể hiện của một lớp, RDF hỗ trợ thuộc tính rdf:type để diễn tả quan hệ ”là thể hiện của”. Ví dụ: tuyên bố

<http://ronaldo.com/foaf.rdf#me> rdf:type foaf:Person .

cho biết Ronaldo là một thể hiện của foaf:Person, một từ vựng của ontology FOAF đại diện cho một con người.

Vì rdf:type là thuộc tính cơ bản của RDF và được dùng thường xuyên trong thực tế, cú pháp Notation 3 và Turtle sử dụng từ "a" để viết tắt cho rdf:type.

Hình 1.3 minh họa một đồ thị RDF nhiều nút.



Hình 1.3 *Ví dụ minh họa một đồ thị RDF nhiều nút*

1.3.2.4 Tài nguyên URI, nút trắng, và giá trị hằng

Nếu một tài nguyên được xác định bởi một URI thì nó được gọi là tài nguyên URI, ngược lại nó là một nút trắng ẩn danh. Vị ngữ trong câu luôn là một tài nguyên URI. Các thuộc tính RDF trên thực tế là các tài nguyên RDF cụ thể hơn. Chúng thuộc lớp rdf:Property được định nghĩa trong RDFS. Vì vậy, các thuộc tính này được sử dụng như vị ngữ trong một tuyên bố.

Hằng giá trị RDF có hai dạng là xâu ký tự đơn giản và giá trị hằng có định kiểu.

Giá trị hằng định kiểu có thẻ kiểu dữ liệu được sử dụng để biểu diễn các giá trị số, ngày, tháng, Boolean, v.v. RDF dùng kiểu dữ liệu XML và cho phép định nghĩa kiểu dữ liệu tùy chỉnh. Ví dụ, trong cú pháp Turtle hằng giá trị định kiểu "22.30"^^xsd:float biểu diễn số thực 22.30.

Xâu ký tự đơn giản có thêm thẻ ngôn ngữ. Ví dụ trong cú pháp Turtle xâu ký tự "Cristiano Ronaldo"@en cho thấy ngôn ngữ của các ký tự đơn giản này là tiếng Anh. Điều này cho phép thêm nhiều xâu ký tự của nhiều ngôn ngữ khác nhau vào đồ thị.

1.3.3 RDFS (RDF SCHEMA)

RDFS (RDF Schema) mở rộng bộ từ vựng RDF Core. Nó chứa một số khái niệm được định nghĩa trước để định nghĩa mới các lớp (chính là các khái niệm) và các thuộc tính của ontology như rdfs:Class, rdfs:Property v.v.

1.3.3.1 Các lớp và các thuộc tính

Trong RDF, về cơ bản mọi tài nguyên có thể được sử dụng như một vị ngữ hoặc một lớp (được chỉ định dùng thuộc tính rdf:type). Ví dụ như:

<http://ronaldo.com/foaf.rdf#me> rdf:type foaf:Person .

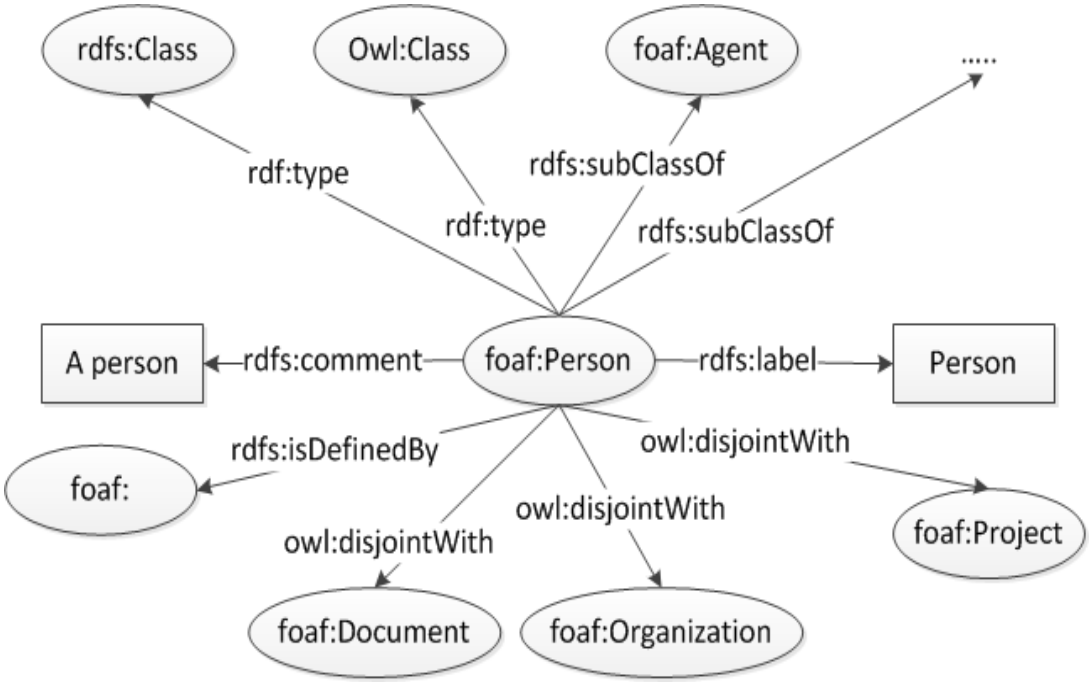
<http://ronaldo.com/> dc:creator <http://ronaldo.com/foaf.rdf#me> .

Tuy nhiên, để có thể hiểu được ngữ nghĩa của foaf:Person và dc:creator, những tài nguyên này phải được mô tả ở đâu đó. Nơi chứa các định nghĩa này chính là ontology mà RDFS (cùng với OWL) là một trong những ngôn ngữ biểu diễn. Trong ví dụ trên foaf:Person là một lớp (hay khái niệm) của ontology FOAF Friends-of-a-Friend [84] được công bố tại http://xmlns.com/foaf/spec/, còn dc:creator là một thuộc tính của bộ từ vựng Dublin Core được định nghĩa tại http://dublincore.org/documents/dcmi-terms/.

Khai báo rdfs:Class cho phép định nghĩa một khái niệm RDF. Nói cách khác một lớp (hay một khái niệm) chính là một thể hiện của rdfs:Class:

```
@prefix vs:    <http://www.w3.org/2003/06/sw-vocab-status/ns#> .                    1
foaf:Person    a                  rdfs:Class,    owl:Class ;                       2
               rdfs:comment       “A person.” ;                                    3
               rdfs:isDefinedBy   “foaf:” ;                                        4
               rdfs:label         “Person” ;                                       5
               rdfs:subClassOf    foaf:Agent , <http://xmlns.com/wordnet/1.6/Agent> ,   6
                                  <http://www.w3.org/2000/10/swap/pim/contact#Person> ,  7
                                  <http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing>,  8
                                  <http://xmlns.com/wordnet/1.6/Person> ;          9
               owl:disjointWith   foaf:Document , foaf:Organization, foaf:Project .   10
               vs:term_Status     “stable” .                                       11
```

Ví dụ trên giải thích ý nghĩa của foaf:Person, nó nói rằng một FOAF Person là một lớp con của các khái niệm khác như foaf:Agent và http://xmlns.com/wordnet/1.6/Agent. Dạng biểu diễn đồ thị của ví dụ trên được thể hiện trong hình 1.4 dưới đây:



Hình 1.4 *Định nghĩa FOAF Person như một phần của bảng từ vựng FOAF*

Bộ từ vựng RDFS chứa một số từ vựng cho phép mô tả chính xác và bổ sung thông tin chi tiết về các khái niệm và thuộc tính. Ví dụ như rdfs:label là nhãn mô tả ngắn và thể hiện rõ ý nghĩa của khái niệm hay thuộc tính. Còn rdfs:comment là chú giải bao quát hơn. Tất cả các giá trị chuỗi ký tự có thể được mô tả trong nhiều ngôn ngữ khác nhau.

1.3.3.2 Miền áp dụng và phạm vi giá trị của các thuộc tính (Domain and Range of Properties)

Cho một thuộc tính xác định p', tập các tuyên bố RDF (s, p', o) có thể được diễn dịch thành quan hệ nhị phân p'(s, o), nó liên kết giá trị o với một chủ ngữ nào đó s. Sử dụng cách biểu diễn như trên, miền Dp' là tập các giá trị có thể của s và phạm vi Rp' là tập các giá trị có thể của o. RDFS cung cấp hai thuộc tính dùng để định nghĩa miền và phạm vi của một thuộc tính RDF. Ví dụ, thuộc tính rdf:type được định nghĩa như sau:

```
rdf:type                                                                    1
        rdf:type          rdf:Property ;                                    2
        rdfs:label        "type" ;                                          3
        rdfs:comment      "The subject is an instance of a class" ;         4
        rdfs:domain       rdfs:Resource ;                                   5
        rdfs:range        rdfs:Class ;                                      6
        rdfs:isDefinedBy  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .   7
```

Ví dụ này chỉ ra rằng những từ vựng RDF Core và RDFS được định nghĩa linh hoạt trong chính RDF. Tuyên bố ở dòng 2 xác định rdf:type là một thể hiện của lớp rdf:Property, nghĩa là nó là một thuộc tính. Miền của rdf:type là bất kỳ tài nguyên RDF nào (dòng 5) và phạm vi là tập của tất cả các lớp (dòng 6).

Những thuộc tính trong RDF nếu không có miền và phạm vi xác định, chúng có thể được dùng với bất kỳ tài nguyên nào mà không cần quan tâm lớp của tài nguyên đó. Đây là một khác biệt lớn với một lược đồ cơ sở dữ liệu, nơi mà thuộc tính được định nghĩa trong ngữ cảnh của một quan hệ cụ thể.

1.3.3.3 Hệ thống kiểu (Type System)

Bên cạnh việc định nghĩa các thể hiện của một lớp bằng thuộc tính rdf:type, RDFS cung cấp một cách thức để định nghĩa phân cấp lớp. RDFS cung cấp một hệ thống định kiểu để mô hình phân cấp lớp theo hướng cụ thể hóa, khái quát hóa. Các lớp có thể định nghĩa như lớp con của lớp khác với thuộc tính rdf:subClassOf. Ví dụ:

```
foaf:Person    rdf:subClassOf    foaf:Agent .
```

Theo luật suy diễn kéo theo RDFS [85], bất kỳ thể hiện nào của foaf:Person cũng là thành viên của foaf:Agent. Các thuộc tính con cũng được xác định theo cách tương tự. Ví dụ để tìm ra tác giả của một bài báo hội nghị, người mà bình phẩm viên cần liên hệ, thì có một dc:creator chuyên dụng hơn được định nghĩa:

```
ex:primaryAuthor    rdf:type              rdf:Property ;
                    rdfs:subPropertyOf    dc:creator .
```

Do hệ quả của luật suy diễn kéo theo RDF-S, nếu một bài báo nào đó có ex:primaryAuthor (tác giả chính) là Johnson, có nghĩa là dc:creator (tạo viên) của bài báo đó là Johnson.

1.3.4 OWL (Web Ontology Language)

OWL (Web Ontology Language) là một ngôn ngữ biểu diễn tri thức hiện đại đã và đang được phát triển dựa trên RDF nhằm cho phép biểu diễn thông tin phân tán trên Web với mức độ biểu đạt cao và suy diễn trên những thông tin đó.

Tùy thuộc vào khả năng diễn tả cần có của một ứng dụng, về cơ bản có ba biến thể khác nhau của OWL [53]:

- OWL Lite
- OWL-DL
- OWL Full

Ngôn ngữ Ontology càng có khả năng diễn đạt thì bộ suy diễn càng phải áp dụng nhiều luật hơn và càng ảnh hưởng tới hiệu năng tính toán. Do đó, sự phân chia nói trên nhằm tạo ra các phiên bản ngôn ngữ phù hợp với đặc thù cụ thể của bài toán ứng dụng.

OWL Lite có hình thức phức tạp thấp nhất. Nó bổ sung một vài tính năng cho RDFS, ví dụ như những ràng buộc đẳng thức/bất đẳng thức cho lớp và cá thể, hoặc những ràng buộc lực lượng cho thuộc tính (nhưng chỉ có giá trị 0 hoặc 1).

OWL-DL được thiết kế để hướng tới khả năng diễn đạt tối đa trong khi vẫn đảm bảo tính đầy đủ và tính giải được của quá trình tính toán (bộ suy diễn sẽ kết thúc trong thời gian nhất định).

OWL Full không đặt ra bất kỳ giới hạn nào đối với thành phần cấu trúc có sẵn của ngôn ngữ (ví dụ, lớp có thể là thể hiện của lớp khác ở cùng một thời điểm, điều này không được phép trong OWL-DL). Nó cung cấp khả năng diễn tả tối đa nhưng không đảm bảo tính giải được.

Tiểu mục này sẽ trình bày tóm tắt những đặc tính quan trọng nhất của OWL và sự khác biệt với RDF/RDFS.

1.3.4.1 Tiên đề và các luật suy diễn kéo theo

Các khuyến nghị W3C về ngữ nghĩa OWL và ngữ nghĩa RDF [86] định nghĩa các cơ chế suy diễn các ontology OWL và RDFS. Các đặc tả cũng bao gồm các tiên đề và các luật mà một bộ suy diễn cần biết để tạo ra chính xác các sự kiện. Tất cả phát biểu định nghĩa sẵn có của RDF Core và RDFS chính là các tiên đề. Ví dụ:

rdf:type rdf:type rdf:Property .

là một tiên đề. Sau đây là một ví dụ về luật suy diễn kéo theo. Cho trước đồ thị RDF có tên là E:

u rdfs:subClassOf x .
v rdf:type u .

với u, v là tham chiếu URI bất kỳ hoặc định danh nút trắng, và x là tham chiếu URI bất kỳ, định danh nút trắng, hoặc chuỗi ký tự. Bộ ba sau có thể được suy ra:

v rdf:type x .

1.3.4.2 Các tính năng của OWL

Các tính năng cơ bản được hỗ trợ bởi cả ba phiên bản OWL là:

1. Các phần tử lược đồ RDF (RDF Schema elements): lớp, cá thể (thể hiện), và thuộc tính; miền và phạm vi của thuộc tính, quan hệ lớp con và thuộc tính con, các kiểu dữ liệu.
2. Đẳng thức/bất đẳng thức (Equality/Inequality): lớp, thuộc tính và cá thể tương đương; các cá thể khác biệt.
3. Đặc tính của thuộc tính (Property characteristics): nghịch đảo, bắc cầu, đối xứng, hàm, quan hệ của thuộc tính chức năng nghịch đảo.
4. Ràng buộc về định lượng của các giá trị của thuộc tính (Restriction on quantification of property values): định lượng với mọi (all values from…), và định lượng tồn tại (some values from…). Lưu ý rằng ràng buộc này được định nghĩa dựa trên một thuộc tính được sử dụng với một lớp cụ thể. Để ràng buộc tổng quát trên phạm vi của một thuộc tính, người ta dùng cấu trúc range của RDFS.
5. Ràng buộc lực lượng (Cardinality restriction): lực lượng có thể bị giới hạn bằng cận trên và cận dưới cũng như bằng một giá trị chính xác. Ví dụ, để chỉ rằng một đội bóng đá có chính xác 11 cầu thủ là hợp lệ.
6. Giao lớp (Class intersection): các lớp mới có thể được định nghĩa như là giao của các lớp khác nhau. Ví dụ, một lớp người vừa là cầu thủ lại vừa là huấn luyện viên có thể được định nghĩa là giao của lớp cầu thủ và lớp huấn luyện viên.

1.3.4.3 Những tính năng bổ sung trong OWL Full và OWL-DL

1. Lớp liệt kê (Enumerated classes): định nghĩa một lớp dựa trên liệt kê các cá thể.
2. Ràng buộc trên giá trị của thuộc tính (Property value restriction): ràng buộc thuộc tính trên một giá trị cụ thể. Ví dụ, lớp cầu thủ Brazil là tất cả những cầu thủ mà thuộc tính quốc gia của họ có giá trị là Brazil.
3. Tính rời nhau của lớp (Disjointness of classes): hai phiên bản OWL trên cho phép tuyên bố tính rời nhau của các lớp
4. Định nghĩa lớp dựa trên tập hợp (Set-based class definition): định nghĩa một lớp dựa trên Tập-kết hợp các lớp khác được xác định bằng các phép hợp, giao, phần bù.

Với việc hỗ trợ tập tính năng phong phú, Ontology OWL có thể biểu diễn tri thức phức tạp khá chính xác. Bộ suy diễn có thể suy ra bộ ba bổ sung dựa trên các luật suy diễn kéo theo đã được định nghĩa trước.

1.4 Tìm kiếm ngữ nghĩa

Tìm kiếm ngữ nghĩa là phương pháp cải thiện độ chính xác tìm kiếm bằng cách hiểu mục đích của người tìm kiếm và ý nghĩa theo bối cảnh của các thuật ngữ tìm kiếm khi chúng xuất hiện trong không gian dữ liệu tìm kiếm, trên mạng hay trong một hệ thống khép kín, để sinh ra các kết quả phù hợp hơn.

Tìm kiếm ngữ nghĩa thể hiện thế mạnh vượt trội của Web ngữ nghĩa trong lĩnh vực tìm kiếm thông tin. Khác với các mô tơ tìm kiếm truyền thống tập trung đếm tần số xuất hiện của từ, các mô tơ tìm kiếm ngữ nghĩa cố gắng hiểu ý nghĩa ẩn tàng bên trong truy vấn của người dùng và cả bên trong các thông tin phản hồi. Dựa vào sự tìm hiểu các công trình [87], luận án nhận thấy tìm kiếm ngữ nghĩa có những dạng thức cơ bản như sau:

- Tìm kiếm dựa trên giao diện người dùng theo ngữ nghĩa: hệ thống tìm kiếm các thông tin theo truy vấn ban đầu, người dùng dựa vào các thông tin này và chọn thông tin bổ sung cho truy vấn ban đầu của mình. Hệ thống dựa vào đó sẽ tìm kiếm hoặc sắp xếp lại các thông tin trả về cho người dùng.
- Tìm kiếm hỏi đáp: hệ thống tìm kiếm các trả lời tương ứng cho một câu hỏi hơn là các tài liệu chứa câu trả lời.
- Truy tìm tài liệu ngôn ngữ có cấu trúc: hệ thống truy tìm thông tin được thể hiện trong các tài liệu ngôn ngữ có cấu trúc ví dụ như sử dụng ngôn ngữ RDF, hoặc sử dụng ngôn ngữ OWL.
- Truy tìm tài liệu ngôn ngữ tự nhiên: hệ thống sử dụng ngôn ngữ tự nhiên để thể hiện truy vấn, và truy tìm các tài liệu được viết bằng các ngôn ngữ tự nhiên. Trong quá trình tìm kiếm, các truy vấn và tài liệu có thể được chú thích ngữ nghĩa. Các tài liệu trả về sẽ được xếp hạng theo độ liên quan với truy vấn.

1.4.1 Các ngôn ngữ truy vấn RDF

Tìm kiếm ngữ nghĩa có thể được thực hiện thông qua phương tiện là các ngôn ngữ truy vấn ngữ nghĩa. Ngôn ngữ truy vấn ngữ nghĩa là ngôn ngữ cung cấp nền tảng cho trích rút thông tin từ đồ thị ngữ nghĩa. Trong một đồ thị ngữ nghĩa thông tin được biểu diễn bởi các đỉnh của đồ thị được liên kết với nhau bởi các cạnh. Cấu trúc của đồ thị được mô tả bằng một ontology trong đó định nghĩa các loại đỉnh, các loại cạnh và các cạnh được liên kết với các đỉnh như thế nào để tạo thành một đồ thị có hướng.

Nghiên cứu về các ngôn ngữ truy vấn RDF đã phân chia chúng thành ba nhóm chính căn cứ vào sự khác biệt về mô hình dữ liệu, tính diễn tả, hỗ trợ thông tin lược đồ và các kiểu truy vấn. Ba nhóm này là:

- SPARQL [51]: ngôn ngữ truy vấn này có nguồn gốc từ ngôn ngữ SquishQL, sau đó phát triển thành RDQL [88] và cuối cùng được mở rộng thành SPARQL. Nhóm ngôn ngữ này xem RDF như là dữ liệu bộ ba mà không quan tâm đến lược đồ hay thông tin về Ontology trừ khi điều đó được nêu rõ trong nguồn RDF.

- RQL [89] và mở rộng của nó SeRQL [90]: nhóm này có điểm chung là hỗ trợ kết hợp truy vấn dữ liệu và lược đồ. Mô hình dữ liệu RDF được sử dụng hơi lệch so với mô hình dữ liệu chuẩn của RDF và RDFS, do đó làm mất đi các chu trình trong phân cấp bao hàm và các yêu cầu về cả miền xác định và miền giá trị định nghĩa cho mỗi thuộc tính. Mặt khác, ngôn ngữ này khá là phức tạp khiến khả năng biểu diễn của nó yếu hơn so với SPARQL.
- TRIPLE [91]: vừa là ngôn ngữ truy vấn vừa là ngôn ngữ luật. TRIPLE không có khả năng phân biệt giữa luật và truy vấn. TRIPLE cũng không tin cậy vì nó cho phép thực hiện các luật không chắc chắn. Các ngữ nghĩa mong muốn phải được chi tiết hóa thành một tập luật đi cùng với truy vấn. TRIPLE không hỗ trợ kiểu dữ liệu.

1.4.2 SPARQL

SPARQL (SPARQL Protocol and RDF Query Language) [51] là một ngôn ngữ truy vấn dữ liệu RDF được phát triển mới bởi nhóm RDF Data Access Working Group và được W3C khuyến cáo chính thức từ năm 2008 do các ưu điểm so với các ngôn ngữ truy vấn khác như Triple, RDQL, RQL, SeRQL v.v. SPARQL giúp truy vấn thông tin từ Ontology nhanh chóng và hiệu quả. SPARQL hỗ trợ hầu hết các tính năng truy vấn cần có như là: hỗ trợ mô hình dữ liệu RDF, tính đóng, tính đầy đủ, tính trực giao, biểu thức đường dẫn, OPTIONAL Path, phép hợp UNION, phép hiệu DIFFERENCE, định lượng, tổng hợp và gom nhóm.

Chính vì vậy SPARQL là một lựa chọn tốt cho các truy vấn ngữ nghĩa. Dưới đây là một số dạng truy vấn SPARQL thường dùng:

1.4.2.1 Truy vấn SELECT…WHERE

Truy vấn dạng này gồm 2 mệnh đề:

- Mệnh đề SELECT chỉ ra những biến cần tìm.
- Mệnh đề WHERE chỉ ra các mẫu đồ thị - là điều kiện cần khớp của các biến.

Ví dụ:

```
select ?uri ?label where {
        ?uri rdf:type BKSport:Stadium.
        ?uri rdfs:label ?label.
        filter(lang(?label)='en')
        ?uri BKSport:hasLocation ?location.
        ?location BKSport:isPartOf BKSport:manchester-city.
        ?uri BKSport:isWellKnown "true"^^xsd:boolean.
}
```

1.4.2.2 Truy vấn ASK

Truy vấn này tương tự truy vấn SELECT…WHERE nhưng có những điểm khác như sau:

- Không cần chỉ ra các biến cần lấy giá trị, chỉ cần chỉ ra các mẫu đồ thị.
- Kết quả trả về là giá trị logic:
    - True: nếu tồn tại lời giải.
    - False: nếu không tồn tại lời giải.

Ví dụ:

```
ask {
BKSport:manchester-city-footballclub rdf:type BKSport:FootballClub.
}
```

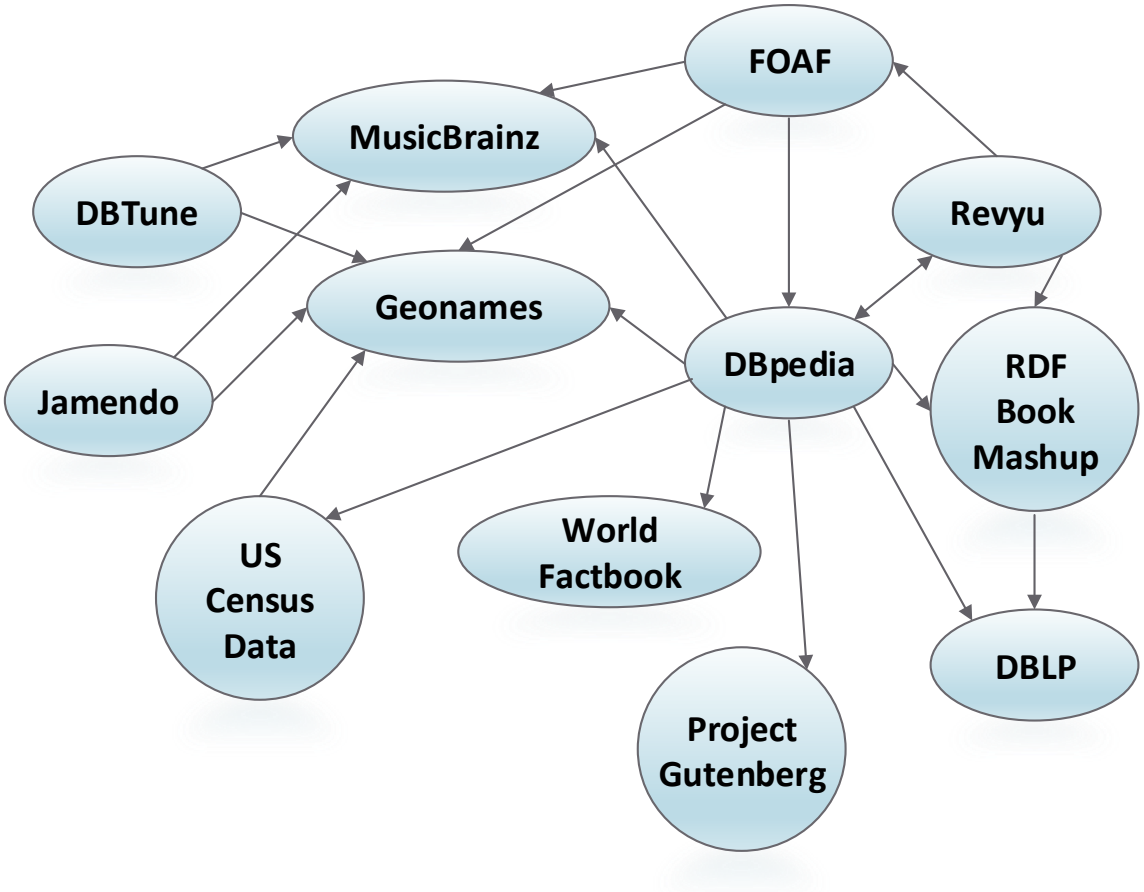
Ngoài ra SPARQL còn hỗ trợ các dạng truy vấn khác như CONSTRUCT, DESCRIBE.

1.5 Kho dữ liệu ngữ nghĩa mở

Công nghệ Web ngữ nghĩa cung cấp một môi trường để tạo và xuất bản dữ liệu có cấu trúc trên Web. Theo Tim Berners-Lee, siêu dữ liệu có thể hữu ích hơn, nếu nó được biểu diễn với các từ vựng chung (tái sử dụng các ontology hiện có) và được kết nối tới các tập dữ liệu khác nhau trên Web (các liên kết giữa các tập dữ liệu). Từ những nhu cầu này, thuật ngữ Dữ Liệu Liên Kết đã được đưa ra bởi Tim Berners-Lee trong ghi chú của ông về kiến trúc Web Dữ Liệu Liên Kết. Thuật ngữ này nói đến trình bày, chia sẻ và kết nối dữ liệu có cấu trúc trên Web ngữ nghĩa. Giá trị và tính hữu dụng của dữ liệu tăng hơn khi nó được kết nối với dữ liệu khác. Khi dữ liệu được công bố trên Web ngữ nghĩa và được kết nối với các tập dữ liệu khác, việc khám phá thông tin có thể được cải thiện. Dữ Liệu Liên Kết là kết quả của một nỗ lực cộng đồng. Dự án cộng đồng Dữ Liệu Mở Liên Kết của nhóm W3C Semantic Web Education and Outreach hướng đến tăng cường Web của Dữ Liệu Liên Kết bằng cách xuất bản các tập dữ liệu mở khác nhau ở định dạng RDF trên Web và bằng cách kết nối chúng tới các nguồn dữ liệu khác nhau.

Một số ví dụ về Dữ Liệu Liên Kết là: DBpedia [92], Faceted DBLP [93], Geonames [94]. DBpedia là một nỗ lực cộng đồng nhắm đến trích xuất thông tin có cấu trúc từ Wikipedia và xuất bản thông tin này trên Web ngữ nghĩa và liên kết các tài nguyên này tới các tập dữ liệu khác nhau. Cơ sở dữ liệu thư mục học DBLP cung cấp siêu dữ liệu về các bài báo khoa học, các hội nghị, các tạp chí và các tác giả. Geonames cung cấp siêu dữ liệu về dữ liệu địa lý (ví dụ tên các địa điểm trong các ngôn ngữ khác nhau, dân số v.v) và vĩ độ/ kinh độ của địa điểm.

Hình 1.5 dưới đây minh họa một phần của dữ liệu liên kết mở trên Web đến ngày 8 tháng 1 năm 2019.



Hình 1.5 *Một phần của Dữ Liệu Liên Kết Mở trên Web, ngày 8 tháng 1 năm 2019 [95]*

Nguyên lý cơ bản của Dữ Liệu Liên Kết

Trong [96], các tác giả đã đưa ra một tập các quy tắc dưới tên gọi "Nguyên Lý về Dữ Liệu Liên Kết" để xuất bản dữ liệu trên Web theo một cách mà tất cả dữ liệu được xuất bản trở thành một bộ phận của một không gian dữ liệu toàn cầu:

- Sử dụng URI để định danh các sự vật (các tài nguyên)
- Sử dụng các HTTP URI để con người và ứng dụng có thể tìm kiếm và tra cứu một URI qua giao thức HTTP.
- Khi một người tra cứu một URI, phải cung cấp được các thông tin hữu ích sử dụng các chuẩn như RDF, SPARQL
- Liên kết với các dữ liệu khác. Mô tả tài nguyên cần chứa các liên kết tới các URI liên quan trong các phát biểu RDF hoặc như các liên kết rdfs:seeAlso hoặc owl:sameAs.

Trong khi đơn vị cơ bản của Web siêu văn bản là các tài liệu HTML kết nối với nhau bởi các siêu liên kết không định kiểu, Dữ Liệu Liên Kết dựa trên các tài liệu chứa dữ liệu ở định dạng RDF. Tuy nhiên, thay vì chỉ đơn giản kết nối các tài liệu đó, Dữ Liệu Liên Kết sử dụng RDF để tạo ra các tuyên bố được định kiểu, liên kết các sự vật riêng lẻ. Kết quả thu được là cái mà chúng ta gọi là Web Dữ Liệu, hiểu một cách chính xác chính là *Web của những sự vật, được mô tả bởi dữ liệu trên Web*.

1.6 Một số lĩnh vực ứng dụng Web ngữ nghĩa

1.6.1 Thương mại điện tử

Lĩnh vực sau của thương mại điện tử có nhiều khả năng hưởng lợi nhờ việc ra đời của công nghệ Web ngữ nghĩa. Quản lý chuỗi cung ứng điện tử (eSCM) là một khái niệm được đưa ra để đáp ứng yêu cầu về khả năng thích ứng và linh hoạt trong một môi trường thương mại điện tử rất năng động, trong đó tập trung vào tích hợp mạng thông qua các liên kết điện tử và cấu trúc dựa trên các quan hệ được kích hoạt công nghệ. Chuỗi cung ứng bản thân nó là một mạng lưới động và phức tạp liên quan đến nhiều nhà cung cấp, nhà sản xuất, các nhà kho, nhà bán lẻ, và khách hàng. Ali Ahmad và cộng sự đề xuất phương pháp luận xây dựng ontology cho lĩnh vực quản lý chuỗi cung ứng trên cơ sở nhận thức rằng ontology sẽ giúp cho việc chia sẻ tri thức và giao tiếp giữa các bên liên quan của hệ thống này trở nên hiệu quả hơn [15].

1.6.2 Chăm sóc sức khỏe và khoa học đời sống (HCLS)

Trong [97], các tác giả cho rằng các hoạt động quản lý tri thức trong chăm sóc sức khỏe tập trung vào việc thu thập và lưu trữ thông tin và hiện nay thiếu khả năng chia sẻ và chuyển giao tri thức giữa các hệ thống và tổ chức để hỗ trợ hiệu quả công việc của người dùng cá nhân. Công nghệ Web ngữ nghĩa có thể cho phép tích hợp thông tin sức khỏe, do đó cung cấp trong suốt cho các tiến trình liên quan đến chăm sóc sức khỏe bao gồm tất cả các thực thể trong và giữa các bệnh viện, cũng như các bên liên quan như hiệu thuốc, nhà cung cấp bảo hiểm, nhà cung cấp dịch vụ chăm sóc sức khỏe, và phòng thí nghiệm lâm sàng. Ứng dụng công nghệ tiên tiến trong khám phá và quản lý tri thức có vai trò quan trọng trong lĩnh vực chăm sóc sức khỏe. Trong [22], tác giả cho rằng Web ngữ nghĩa là khung làm việc phù hợp cho bài toán quản lý tri thức quy mô lớn và phân tán. Để ứng dụng hiệu quả công nghệ này cần vượt qua những thách thức như là phát triển một phương pháp biểu diễn tri thức trực quan nhất quán có cơ sở vững chắc cho những nghiệp vụ chính. Dumontier đề xuất sử dụng các thuật ngữ trong ontology hình thức để biểu diễn các mô tả tri thức và làm tăng liên tác ngữ nghĩa giữa các lĩnh vực con.

1.6.3 Chính phủ điện tử

Những nghiên cứu ứng dụng Web ngữ nghĩa trong lĩnh vực chính phủ điện tử đã bắt đầu từ những năm 2000. Đối với người dùng của các hệ thống này việc tiếp cận và sử dụng số lượng lớn và phức tạp các tài nguyên thông tin như các file, các liên kết, các dịch vụ … là vẫn còn trở ngại. Nghiên cứu của [10] đầu tiên xác định những rào cản ở góc độ ngữ nghĩa của các hệ thống chính phủ điện tử thông thường như trải nghiệm không thỏa mãn của người dùng, thiếu tính liên tác do sự không khớp về ngữ nghĩa của dữ liệu trao đổi, quản lý tài liệu kém do tìm kiếm thông tin không hiệu quả… Klischewski đã lựa chọn sử dụng ontology để biểu diễn cấu trúc ngữ nghĩa của các tài nguyên thông tin. Từ đó tạo ra các mô tả mà máy tính có thể hiểu được

về các thông tin có tính đến ngữ cảnh người dùng. Hệ thống qua đó có thể quyết định việc hiển thị thông tin phù hợp với từng cá nhân. Nghiên cứu cũng chỉ ra các bài toán mà công nghệ Web ngữ nghĩa cần được tiếp tục ứng dụng để giải quyết như về chi phí và lợi nhuận của tổ chức, sự tham gia đóng góp của chuyên gia, tích hợp công nghệ...

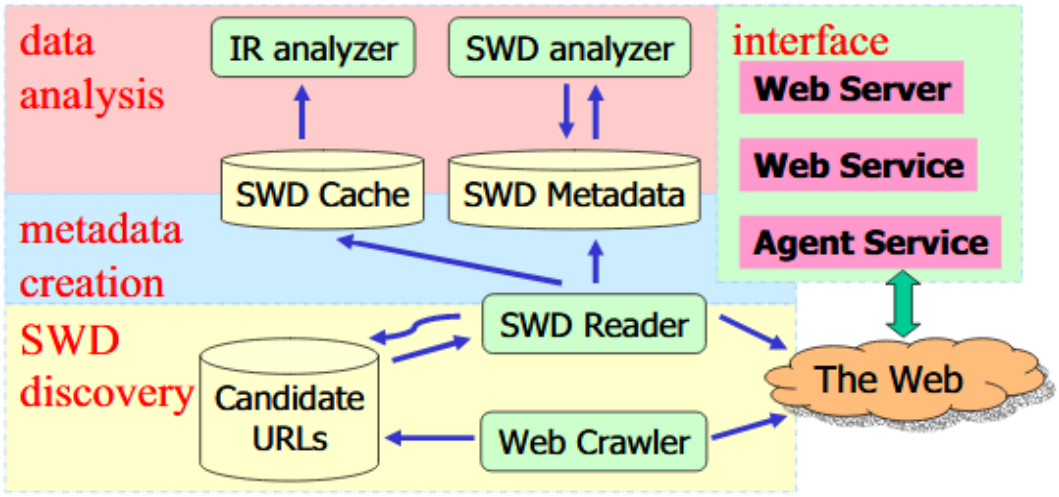
1.6.4 E-Learning

Web ngữ nghĩa là một nền tảng rất phù hợp cho việc thực hiện một hệ thống e-Learning hoàn chỉnh, vì nó đáp ứng được các yêu cầu học đúng lúc và đúng kiến thức. Điều này đã được giải thích trong nghiên cứu của [16] như sau: ontology giúp mô tả và tập hợp được các tài liệu học tập phân tán trên Web mà phù hợp với từng cá nhân người học. Trước đó vào năm 2001, Stojanovic , Staab và Studer đã nêu ra bài toán mà Web ngữ nghĩa có thể trợ giúp e-Learning như sau: người dùng cần tìm những tài liệu học tập mong muốn, hệ thống cung cấp thông tin một cách chủ động để tạo ra một môi trường học tập năng động, tri thức cần được cung cấp dưới nhiều hình thức khác nhau, tạo ra các tác tử đại diện cho mỗi người dùng có khả năng giao tiếp với các tác tử khác để có được tài liệu [98].

1.7 Một số nghiên cứu Web ngữ nghĩa tiêu biểu

1.7.1 Swoogle

Web ngữ nghĩa là một lĩnh vực nghiên cứu ngày càng phát triển và có ứng dụng rộng khắp, trên nhiều lĩnh vực: y tế, nông nghiệp, truyền thông, thương mại điện tử, quản lý tri thức… Cũng vì thế các ontology và các dữ liệu ngữ nghĩa ngày càng phong phú trên Web. Làm sao có thể tìm kiếm được các ontology và dữ liệu ngữ nghĩa phù hợp, từ đó khai thác được chúng đối với những người làm việc trong lĩnh vực Web ngữ nghĩa có vai trò quan trọng, ví dụ như tái sử dụng ontology hay tích hợp dữ liệu ngữ nghĩa. Dự án nghiên cứu phát triển máy tìm kiếm ontology và dữ liệu ngữ nghĩa đã được Li Ding cùng các cộng sự thực hiện từ năm 2004 [7]. Swoogle là sản phẩm của dự án nói trên đã đem lại nhiều tính năng hữu ích. Ngoài việc cho phép tìm kiếm theo từ khóa, hệ thống này còn có thể thực hiện tìm kiếm theo các ràng buộc và thuộc tính của lớp, làm nổi bật các thuộc tính cấu trúc thú vị như Web ngữ nghĩa được kết nối như thế nào, ontology được tham chiếu như thế nào, và một ontology được sửa đổi từ bên ngoài như thế nào. Hình 1.6 dưới đây minh họa kiến trúc của Swoogle.



Hình 1.6 *Kiến trúc của Swoogle [7]*

Bốn thành phần chính của kiến trúc Swoogle là (1) khám phá dữ liệu ngữ nghĩa, (2) tạo siêu dữ liệu, (3) phân tích dữ liệu, và (4) giao diện. Kiến trúc của Swoogle tập trung vào dữ liệu và có thể mở rộng được. Bốn thành phần trên làm việc một cách độc lập và tương tác với nhau thông qua một cơ sở dữ liệu mà chứa siêu dữ liệu về dữ liệu ngữ nghĩa.

Thành phần khám phá dữ liệu ngữ nghĩa tìm kiếm các dữ liệu ngữ nghĩa tiềm năng trên Web và luôn cập nhật các dữ liệu ngữ nghĩa đó. Nó gồm hai thành phần chính, cụ thể là Google Crawler và Focused Crawler. Google Crawler sử dụng dịch vụ Web của Google để thu thập các URL, với sự tập trung vào các mở rộng kiểu như ".rdf", ".owl", ".daml", và ".n3". Trong khi đó, Focused Crawler thu thập các tài liệu bên trong một Website đang tồn tại. Để giảm độ phức tạp tìm kiếm và tăng độ chính xác, các Heuristic đơn giản như ràng buộc mở rộng và ràng buộc tập trung được sử dụng để lọc ra các tài liệu được cho là không thích hợp.

Thành phần tạo siêu dữ liệu gồm một bản sao của dữ liệu Web ngữ nghĩa, và siêu dữ liệu về các dữ liệu Web ngữ nghĩa được sinh khách quan ở cả mức cú pháp và mức ngữ nghĩa. Siêu dữ liệu về dữ liệu ngữ nghĩa được thu thập để làm cho phép tìm kiếm về dữ liệu Web ngữ nghĩa trở nên có hiệu lực và có hiệu quả hơn. Swoogle nhận ra ba loại siêu dữ liệu: (1) siêu dữ liệu cơ bản, (2) các quan hệ, và (3) các kết quả phân tích như phân loại các ontology hay cơ sở dữ liệu Web ngữ nghĩa, xếp hạng dữ liệu Web ngữ nghĩa.

Thành phần phân tích dữ liệu sử dụng các dữ liệu ngữ nghĩa được lưu trữ và siêu dữ liệu được tạo ra để có được các báo cáo phân tích, chẳng hạn như sự phân loại các ontology Web ngữ nghĩa và các cơ sở dữ liệu Web ngữ nghĩa, hạng của các dữ liệu Web ngữ nghĩa, và chỉ số tìm kiếm thông tin của các dữ liệu Web ngữ nghĩa.

Thành phần giao diện tập trung vào cung cấp dịch vụ dữ liệu tới cộng đồng Web ngữ nghĩa. Một giao diện Web đã được thực hiện tại http://www.swoogle.umbc.edu.

1.7.2 Dự án ARTEMIS

Các hệ thống thông tin sức khỏe thường phục vụ riêng cho các tổ chức y tế khác nhau, do đó hạn chế trong việc trao đổi dữ liệu cũng như truy nhập các tính năng của nhau. Cải thiện tính liên tác giữa các hệ thống trên là cần thiết. ARTEMIS [14] là một dự án nghiên cứu với mục tiêu giải quyết vấn đề tính liên tác ở cả mức ngữ nghĩa cũng như mức chức năng. Tính liên tác chức năng được thực hiện thông qua kiến trúc hướng dịch vụ, tính liên tác ngữ nghĩa được tạo ra nhờ các chú thích ngữ nghĩa về các dịch vụ Web nói trên. Kết quả là các dịch vụ Web ngữ nghĩa được tạo ra. Hệ thống ARTEMIS có kiến trúc mạng ngang hàng trong đó các Viện chăm sóc sức khỏe tham gia với vai trò là các phần tử. Mỗi phần tử ngang hàng cung cấp giao diện tới các hệ thống thông tin chăm sóc sức khỏe để cho phép chúng khám phá và sử dụng các dịch vụ Web cung cấp bởi các phần tử khác. Ví dụ như truy cập vào hồ sơ chăm sóc sức khỏe của bệnh nhân, tiếp nhận bệnh nhân, hay hệ thống thông tin phòng thí nghiệm. Các dịch vụ Web ngữ nghĩa có đặc thù là có thể được triệu gọi linh hoạt dựa trên ý nghĩa tính năng của chúng. Ontology giúp xây dựng dịch vụ ánh xạ giữa các dạng biểu diễn dữ liệu khác nhau giữa các tổ chức.

1.7.3 Dartgrid

Trong bất kỳ một miền lĩnh vực nào từ giáo dục, y tế, tài chính, thương mại điện tử, khoa học đều có nhiều CSDL quan hệ được phát triển bởi các quốc gia, tổ chức, cá nhân. Điều đó dẫn đến tính phi thuần nhất của các CSDL này. Mục đích đầu tiên của việc tích hợp các CSDL trên là khai thác sử dụng được nguồn dữ liệu toàn thể đầy đủ. Người dùng cần một giao diện truy vấn dữ liệu thống nhất giúp tạo ra cảm giác như làm việc với một CSDL duy nhất, từ đó họ sẽ dễ dàng làm quen và sử dụng hệ thống thay vì làm việc với nhiều giao diện khác nhau. Các tiếp cận truyền thống gặp những khó khăn nhất định đến từ tính đa dạng trong thiết kế lược đồ quan hệ của các nguồn dữ liệu. Dự án DartGrid [13] được thành lập để giải quyết yêu cầu nói trên với giải pháp kỹ thuật và tiếp cận được lựa chọn là kết hợp Web ngữ nghĩa và tính toán lưới. Web ngữ nghĩa được ứng dụng để tạo ra mức dịch vụ ngữ nghĩa mới, ở đó các lược đồ quan hệ được điều phối và truy vấn ngữ nghĩa được xử lý. Giao diện truy vấn ngữ nghĩa dựa trên ontology được xây dựng. Các thành phần chính của DartGrid là Ontology Service, Semantic Registration Service, Semantic Query Service, Search Service. Ontology Service bộc lộ các ontology được chia sẻ, Semantic Registration Service duy trì thông tin ánh xạ ngữ nghĩa. Semantic Query Service xử lý những truy vấn ngữ nghĩa SPARQL. Search Service hỗ trợ tìm

kiếm toàn văn bản trong tất cả CSDL. Một số tính năng kỹ thuật nổi bật là công cụ ánh xạ ngữ nghĩa trực quan hóa, viết lại truy vấn SPARQL với nhiều khả năng suy luận bổ sung, giao diện người dùng truy vấn ngữ nghĩa dựa trên ontology, công cụ tìm kiếm dựa trên ontology với xếp hạng khái niệm và điều hướng ngữ nghĩa.

1.7.4 Kho nội dung Web ngữ nghĩa cho nghiên cứu lâm sàng

Các tiến bộ trong y tế dẫn tới sự ra đời của rất nhiều cơ sở dữ liệu lớn cho các chuyên ngành riêng. Các dữ liệu được lưu trữ riêng rẽ trong các cơ sở dữ liệu trên cùng với sự nhập nhằng và không thống nhất về thuật ngữ gây khó khăn trong việc tích hợp, và cản trở đổi mới trong nghiên cứu lâm sàng và tin sinh học. Dự án nghiên cứu tại bệnh viện Cleveland [8] có mục tiêu là cải thiện khả năng của bệnh viện bằng cách sử dụng dữ liệu bệnh nhân để sinh ra tri thức mới, cải thiện chăm sóc bệnh nhân trong tương lai thông qua nghiên cứu lâm sàng chiều dọc và tiếp cận Web ngữ nghĩa đã được lựa chọn để tạo ra một kiến trúc tích hợp cho hệ thống.

Kho chứa nội dung thống nhất SemanticDB về dữ liệu bệnh nhân được xây dựng thông qua một phương pháp thu thập dữ liệu, quản lý tài liệu, và biểu diễn tri thức. Nhóm nghiên cứu cũng phát triển ứng dụng để chuyển đổi tự động dữ liệu về RDF. Dữ liệu ngữ nghĩa có thể được biến đổi và lưu trữ trong CSDL MySQL. Kho nội dung này là kênh thông tin chính cho toàn bộ các ứng dụng cung cấp các tính năng tìm kiếm, tổng hợp, tóm tắt thông tin bệnh nhân.

Cơ chế suy diễn ra các tri thức mới, và một hệ chuyên gia hỏi đáp về các bệnh nhồi máu cơ tim cũng được phát triển. Lợi ích chính của sử dụng công nghệ Web ngữ nghĩa là sử dụng thuật ngữ địa phương quen thuộc, hỗ trợ phần mở rộng mô hình hóa không dự kiến trước, hỗ trợ tự động hóa cao, tích hợp có độ chính xác cao và ánh xạ với các hệ thống ngoài và các thuật ngữ, hỗ trợ trả lời chính xác các truy vấn có nghĩa.

1.7.5 Ứng dụng Web ngữ nghĩa trong lĩnh vực nông nghiệp của tổ chức nông-lương thực Liên hiệp quốc (FAO)

Một nhiệm vụ trọng tâm của tổ chức lương thực và nông nghiệp của Liên Hiệp Quốc (Food and Agriculture Organization of the United Nation) là đưa thông tin tới những người cần chúng. Hoạt động này gồm bốn lĩnh vực chính: (1) đưa thông tin vào tầm tay của người dùng, (2) chia sẻ kinh nghiệm về chính sách, (3) tạo ra một nơi gặp gỡ cho các quốc gia, (4) và đưa tri thức vào thực tế. Tuy nhiên các tài nguyên thông tin nông nghiệp có đặc tính phân tán khác nhau, khác biệt về khuôn dạng, và quan trọng nhất là mức độ bao phủ chuyên sâu là khác nhau. Nhóm nghiên cứu của Margherita Sini, Gauri Salokhe và các cộng sự [12] [9] nghiên cứu sử dụng Web ngữ nghĩa nhằm làm tốt hơn các mục tiêu trên.

Ontology AgRIS được xây dựng để bao gồm các khái niệm, từ vựng cần thiết để mô tả các nguồn tài nguyên thông tin nông nghiệp, cũng như các tài liệu (ví dụ tổ chức, loại tài nguyên, các loại chủ đề, tiêu đề tài liệu, người viết, nhà xuất bản…) Ontology này giúp giải quyết trở ngại gây ra do sự không thuần nhất về ngữ nghĩa giữa các nguồn dữ liệu. Ngoài ra, nó còn được dùng cùng với bách khoa thư AGROVOC để mở rộng truy vấn tìm kiếm. Một cổng thông tin được xây dựng cho phép người dùng tra cứu và tìm kiếm các bài báo trong tạp chí Lương thực, Dinh dưỡng và Nông nghiệp (FNA) bao trùm nhiều chủ đề khác nhau. Các bài báo này đều có các metadata mô tả sử dụng AGRIS do đó cho phép thực hiện tìm kiếm ngữ nghĩa, tìm kiếm chính xác theo từ đồng nghĩa.

1.8 Website và cổng thông tin tin tức có ngữ nghĩa

Hiện nay, hầu hết các Website đều lưu trữ dữ liệu trong các Hệ Quản Trị Cơ Sở Dữ Liệu (RDBMS) do các ưu điểm đã được chứng minh của CSDL về khả năng mở rộng, lưu trữ hiệu quả, tối ưu hóa việc thực thi các câu truy vấn, độ an toàn. Tuy nhiên, các CSDL quan hệ (RDB) thường là tách biệt nhau, không thống nhất về lược đồ, thuật ngữ, định danh và mức độ chi tiết của sự biểu diễn dữ liệu. Vấn đề này đang được các nhà khoa học quan tâm và mong muốn tìm ra giải pháp để có thể tái sử dụng và tích hợp nhiều nguồn dữ liệu quý giá và khổng lồ của Web. Để giải quyết vấn đề nêu trên, nhiều nhà khoa học cho rằng có thể sử dụng kỹ thuật RDF và

Ontology của Web ngữ nghĩa để đem đến một nền tảng cho việc tích hợp và công khai tất cả các nguồn dữ liệu đó một cách tự động và trong suốt trên Web.

Cổng thông tin có thể được hiểu như là một điểm truy cập cho việc trình bày, trao đổi, thu thập thông tin từ nhiều nguồn khác nhau trên Internet trong một site duy nhất phục vụ một cộng đồng cụ thể. Trong nghiên cứu [23], Hyvönen phân loại cổng thông tin thành ba loại chính. Loại thứ nhất, cổng thông tin dịch vụ tập hợp một tập lớn các dịch vụ lại với nhau. Trong khi đó, cổng thông tin cộng đồng hành động như nơi gặp gỡ ảo của cộng đồng, và cổng thông tin hướng thông tin thì hoạt động như một kho chứa dữ liệu.

Cổng thông tin hiện nay cho thấy những giới hạn nghiêm trọng liên quan đến các tiện ích cho tìm kiếm, truy cập, rút trích, diễn dịch và xử lý thông tin. Hướng áp dụng các kỹ thuật Web ngữ nghĩa trong xây dựng các cổng thông tin có tiềm năng vượt qua những hạn chế trên. Mặt khác, cũng cần các cổng thông tin ngữ nghĩa có khả năng xuất bản nhiều nội dung Web ngữ nghĩa. Dưới đây là các khái niệm về cổng thông tin ngữ nghĩa được đưa ra từ các góc nhìn khác nhau.

Tác giả Abrahams [99] đưa ra khái niệm cổng thông tin ngữ nghĩa là một tập hợp các tài nguyên dựa trên ontology với các từ khóa tìm kiếm. Việc tìm kiếm tài nguyên trong cổng thông tin ngữ nghĩa thường dựa trên khai thác cấu trúc ontology nêu trên.

Trong [100] của Holger Lausen và các cộng sự, cổng thông tin ngữ nghĩa được định nghĩa là một Website cung cấp thông tin và trao đổi các tiện ích cho một cộng đồng có cùng mối quan tâm dựa trên việc sử dụng công nghệ Web ngữ nghĩa.

Theo Hyvönen [23], cổng thông tin ngữ nghĩa dựa trên các chuẩn Web ngữ nghĩa. Trong đó, Web ngữ nghĩa bao gồm metadata, ontology, và các luật để biểu diễn có cấu trúc, các tính năng mở rộng cho thiết kế các cổng thông tin truyền thống.

Việc áp dụng Web ngữ nghĩa vào cổng thông tin đem lại lợi ích cho nhiều đối tượng khác nhau:

- Đối với người sử dụng, hệ thống này cung cấp cho người sử dụng một cái nhìn tổng quát tới những nội dung phân tán và phi thuần nhất, tự động tổng hợp thông tin [101], tìm kiếm ngữ nghĩa theo các metadata giúp cho việc tìm kiếm chính xác. Reynolds và Shabajee [101] giải thích sự ưu việt của tính năng tìm kiếm này là khả năng biểu diễn ý nghĩa của câu hỏi dựa trên một tập từ vựng được kiểm soát (ontology) và trả về kết quả phù hợp. Một số lợi ích khác là hiển thị các ngữ nghĩa và khuyến nghị nội dung cho người sử dụng, cung cấp các dịch vụ thông minh khác như cá nhân hóa giao diện [102], trực quan hóa ngữ nghĩa và khám phá tri thức.
- Đối với các nhà xuất bản nội dung, cổng thông tin có ngữ nghĩa cho phép tạo nội dung phân tán, duy trì liên kết tự động dựa vào metadata và ontology, tạo ra kênh xuất bản thông tin chia sẻ để giảm chi phí, bổ sung ngữ nghĩa cho các loại thông tin khác, tăng khả năng tái sử dụng nội dung. Ví dụ, các cộng đồng quan tâm có thể chia sẻ truy cập tới cùng thông tin cơ sở trong khi sử dụng cấu trúc duyệt, phương tiện tìm kiếm và định dạng trình bày khác nhau.
- Các nhà phát triển có thể sử dụng ontology trong việc mô hình hóa cấu trúc của cổng thông tin. Điều này giúp cổng thông tin có khả năng hỗ trợ trao đổi dữ liệu trong một cộng đồng chuyên môn và dễ dàng xử lý tự động thông tin.

Các tiểu mục tiếp theo trình bày một số dự án nghiên cứu về cổng thông tin ngữ nghĩa.

1.8.1 Dự án SWEPT

Nghiên cứu về chủ đề này đã thu hút được sự quan tâm nhất định. Để người sử dụng có thể tìm kiếm và lựa chọn khách sạn phù hợp với các thuộc tính khác nhau, nhóm các tác giả [24] đã phát triển một cổng thông tin điện tử Web ngữ nghĩa về du lịch được đặt tên là SWEPT (Semantic Web E-Portal for Tourism) để tìm kiếm và rút trích thông tin về khách sạn ở Pakistan. Ontology MyHotel được thiết kế để chứa các thông tin trên, đồng thời hỗ trợ các từ đồng nghĩa. Cổng thông tin cho phép người dùng có thể đặt các câu truy vấn bằng ngôn ngữ tự nhiên và trả về kết quả thích hợp từ ontology.

1.8.2 Dự án ARKive

ARKive [101] [103] là một dự án đã chứng tỏ rõ ràng cho lợi thế phi tập trung của cổng thông tin ngữ nghĩa. Cổng thông tin này xuất bản các thực thể đa phương tiện miêu tả các loài có nguy cơ tuyệt chủng. Dự án đã nhận thấy rằng các cộng đồng người dùng có mối quan tâm khác nhau cần được duyệt cổng theo nhiều cách khác nhau, tìm kiếm thông tin theo các tiêu chí khác nhau. Do đó, thông tin cần được trình bày trên các giao diện tùy biến theo nhu cầu của họ. Giải pháp của nhóm nghiên cứu được đưa ra là sử dụng ontology làm cấu trúc xương sống cho các tài nguyên trong cổng thông tin ARKive. Sau đó, các cộng đồng người dùng có thể bổ sung thêm phân loại riêng, chú thích, giao diện duyệt phù hợp với nhu cầu của họ. Ngoài ra, dữ liệu của ARKive cũng dễ dàng tích hợp với dữ liệu từ các cổng thông tin khác.

1.8.3 Cổng thông tin Esperonto

Cổng thông tin Esperonto [25] là nền tảng cho dự án EU Esperonto. Nó được sinh từ cổng thông tin tri thức ODESeW được phát triển bởi một nhóm nghiên cứu tại đại học Politécnica de Madrid. Cổng thông tin Esperonto sử dụng 5 ontology lĩnh vực cụ thể là Project ontology, Meeting ontology, Documentation ontology, Organization ontology, và Person ontology. Lược đồ ontology và những thể hiện có thể được thay đổi bởi nhà quản trị và các thành viên đã đăng ký. Người sử dụng cổng thông tin được phân loại thành nhà quản trị, người sử dụng khách, thành viên. Mục thông tin mới được tạo ra sẽ được tự động công bố cho bất kỳ người nào sử dụng cổng thông tin. Ba mức truy cập trong cổng thông tin Esperonto là tìm kiếm dựa trên từ khóa, duyệt ontology và truy tìm tất cả thể hiện cho khái niệm đó và các khái niệm con của nó ở mỗi bước duyệt, và tìm kiếm dựa trên ontology. Điểm mạnh của cổng thông tin Esperonto là các tiện ích quản lý ontology dựa trên WebODE [80]. Tuy vậy, giao diện người dùng trong cổng thông tin Esperonto không thân thiện cho người sử dụng, tính năng xử lý và truy cập thông tin vẫn còn một số hạn chế. Thêm vào đó cổng không cung cấp các chức năng cá nhân hóa.

1.8.4 Mondeca ITM

Mondeca ITM (Intelligent Topic Manager) [26] là một nền tảng phát triển và công cụ cho các hệ thống quản lý tri thức và thu thập tri thức tự động dựa trên công nghệ Web ngữ nghĩa, ontology và xử lý ngôn ngữ học. Nó được tạo ra bởi Mondeca – một nhà cung cấp phần mềm cho thị trường tổ chức tài liệu và quản lý tri thức.

ITM sử dụng kỹ thuật biểu diễn ontology Topic Map để mô hình hóa tri thức và nội dung trong cổng thông tin. Nó sử dụng thêm một ontology biểu diễn bằng OWL để mô tả dữ liệu được quản lý. Hệ thống cung cấp các tính năng quản lý và soạn thảo ontology đơn giản nhưng không hỗ trợ công cụ suy diễn. Các nhà phát triển có thể sử dụng các hàm API của Mondeca ITM với đầu ra ở định dạng XML, nhưng chưa thể hưởng lợi từ các dịch vụ Web hay dịch vụ Web ngữ nghĩa như ở hệ thống khác.

Hệ thống hỗ trợ ba chức năng truy cập thông tin: duyệt cấu trúc, tìm kiếm qua từ khóa, và tìm kiếm ngữ nghĩa. Ba chức năng trên giúp người dùng tìm kiếm và duyệt thông tin một cách trực quan. Tuy nhiên việc hỗ trợ cá nhân hóa người dùng không thiết lập được quyền của họ. Mondeca ITM dùng quá nhiều hệ thống tri thức khiến cho hệ thống này trở nên phức tạp. Ưu điểm nổi bật của Mondeca ITM là chọn các khái niệm và tìm các khái niệm ontology được khai thác tốt vào quá trình truy cập thông tin, tạo và bảo trì thông tin.

1.9 Ứng dụng Web ngữ nghĩa trong lĩnh vực thể thao

Đã có một vài nghiên cứu ứng dụng công nghệ Web ngữ nghĩa trong lĩnh vực thể thao nhưng chưa nhiều.

Ứng dụng Web ngữ nghĩa trong tổng hợp tin tức, tìm kiếm và xuất bản là một lĩnh vực nghiên cứu đầy hứa hẹn. BBC là hãng truyền thông dịch vụ công đầu tiên đi theo xu hướng này. Hãng này đã xây dựng Website Giải vô địch bóng đá thế giới FIFA World Cup 2010 theo kiến trúc xuất bản ngữ nghĩa động [104].

Một số nghiên cứu khác chú thích ngữ nghĩa hình ảnh, đoạn phim quay về cuộc thi đấu thể thao. Falcon-S [41] thu thập trên Web để lấy những hình ảnh thuộc lĩnh vực bóng đá, phân tích bối cảnh của những hình ảnh đó, lập chỉ mục chúng theo đối tượng đội bóng, cầu thủ v.v mà có trong cơ sở tri thức. Nhóm tác giả [105] giới thiệu một khung chung cho chú thích ngữ nghĩa, lập chỉ mục và tìm kiếm các trận thi đấu thể thao dựa trên văn bản web-casting và video thể thao phát quảng bá. Trong khung này, họ đã đề xuất một tiếp cận mới cho phân tích văn bản, phân tích video, căn chỉnh văn bản/video và tìm kiếm được cá nhân hóa.

Một số tổ chức đã xây dựng Ontology về thể thao. Hãng truyền thông BBC [106] [107] đã có những nghiên cứu đầu tiên về sử dụng Ontology và kho dữ liệu ngữ nghĩa Dbpedia tích hợp CSDL thuộc về nhiều lĩnh vực. Muthu lakshmi và Uma [108] đã xây dựng một Ontology giáo dục trực tuyến cung cấp các ngữ nghĩa mong muốn cho người học về lĩnh vực thể thao.

1.10 Tiếp cận Web ngữ nghĩa xây dựng hệ thống tin tức thể thao

Như đã trình bày ở trên, công nghệ Web ngữ nghĩa đem lại nhiều lợi ích khi áp dụng vào các hệ thống thông tin, phần mềm trong nhiều lĩnh vực khác nhau. Chức năng cơ bản nhất trong các hệ thống thông tin, các cổng thông tin là tra cứu có thể được cải thiện. Nhiều nghiên cứu ứng dụng Web ngữ nghĩa phát triển tính năng tra cứu, đánh dấu thông tin (bookmark) [11] và mở rộng tìm kiếm dựa trên thuật ngữ của ontology [12] [9]. Web ngữ nghĩa giúp nâng cao chất lượng xử lý thông tin như chẩn đoán, tìm kiếm thông minh dựa trên suy diễn ngữ nghĩa [8] [12].

Ngữ nghĩa mô tả về dịch vụ Web giúp việc xử lý được tự động hóa. Các khái niệm, thông tin, tri thức có cấu trúc phức tạp và chưa có sự thống nhất về cách thức biểu diễn cũng có thể được mô hình hóa sử dụng ontology [9]. Việc tích hợp sử dụng ontology giúp giảm thiểu và giải quyết vấn đề nhập nhằng về thuật ngữ giữa các CSDL, các hệ thống con trong một hệ thống tổng thể [8].

Khảo sát cũng cho thấy những ứng dụng của Web ngữ nghĩa trong lĩnh vực thể thao nói chung và tin tức thể thao còn chưa được quan tâm. Với những kết quả nghiên cứu ứng dụng Web ngữ nghĩa đã công bố, luận án lựa chọn Web ngữ nghĩa là tiếp cận chủ đạo trong việc giải quyết những hạn chế trong tìm kiếm, sắp xếp, trực quan hóa thông tin nhằm đạt được mục tiêu nghiên cứu chung.

Tư tưởng chủ đạo của tiếp cận là như sau. Đầu tiên với mỗi đơn vị thông tin cơ bản của hệ thống là tin tức, cần tạo ra một tầng ngữ nghĩa mới mô tả những gì mà người dùng quan tâm trong tin tức đó. Thay vì lựa chọn mô hình biểu diễn thông tin truyền thống, luận án dựa trên mô hình biểu diễn tin tức thể thao có ngữ nghĩa. Điều đó dẫn đến việc nghiên cứu xây dựng một ontology về thể thao.

Đặc thù của các hệ thống tổng hợp tin tức là phải làm việc với một số lượng lớn các tin tức. Việc sử dụng các công cụ biên tập chú thích ngữ nghĩa thủ công chắc chắn chưa phải là giải pháp toàn diện. Bài toán quan trọng đầu tiên là nghiên cứu các phương pháp, kỹ thuật để sinh ra chú thích ngữ nghĩa cho một số lượng lớn tin tức.

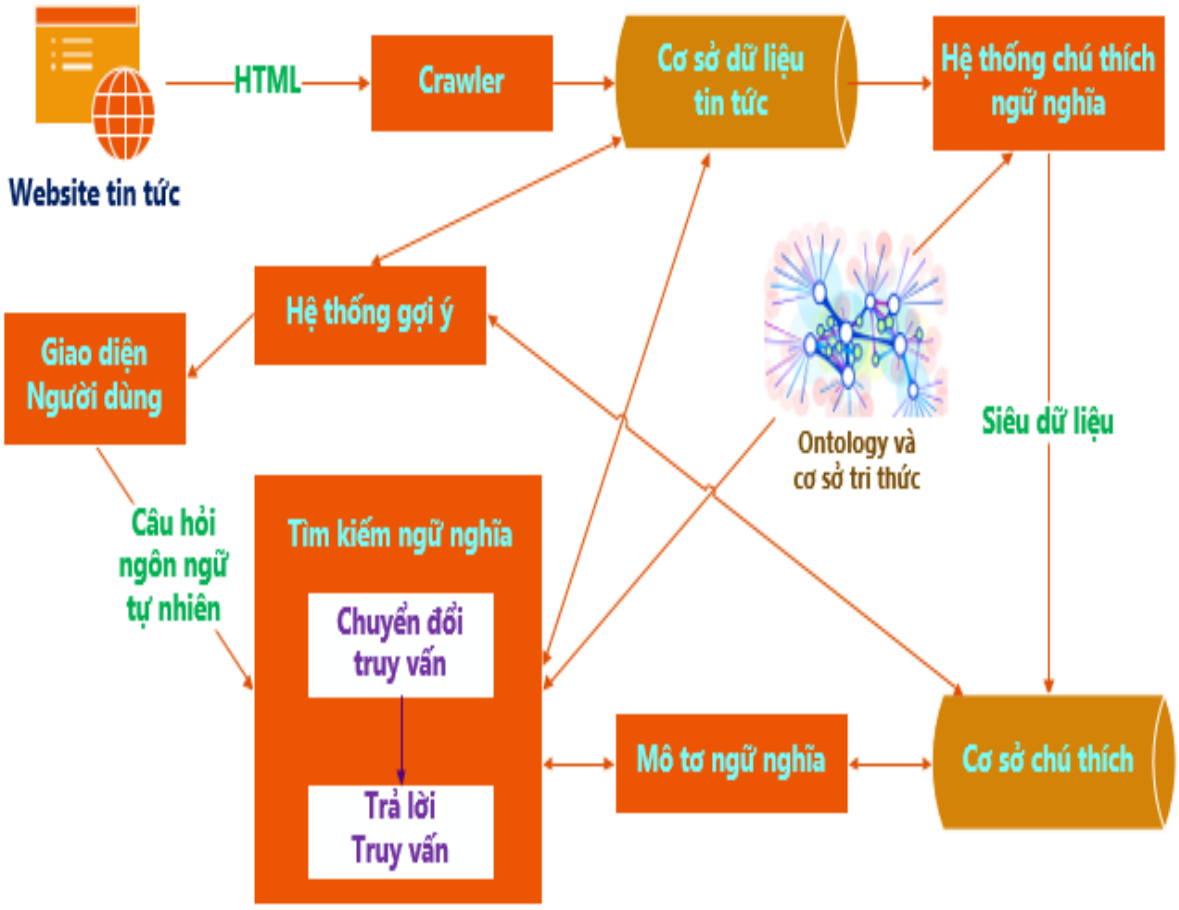
Sau khi đã có được các chú thích ngữ nghĩa cho tin tức, bài toán nghiên cứu tiếp theo là khai thác chúng như thế nào để tạo một hệ thống tổng hợp tin tức như mục tiêu mà luận án hướng tới. Luận án đặt trọng tâm vào việc cải tiến tính năng tìm kiếm và gợi ý tin tức, ứng dụng công nghệ ngữ nghĩa.

Như vậy, công nghệ ngữ nghĩa là công nghệ nền tảng và xuyên suốt trong ba bài toán nghiên cứu chính của luận án: sinh chú thích ngữ nghĩa, tìm kiếm ngữ nghĩa tin tức, gợi ý tin tức dựa trên ngữ nghĩa. Ở mục tiếp theo luận án đề xuất một mô hình kiến trúc cho hệ thống tổng hợp tin tức thể thao. Kiến trúc tổng thể này sẽ minh họa tiếp cận Web ngữ nghĩa được thể hiện trong các thành phần của hệ thống như thế nào. Đồng thời, nó cũng làm rõ vai trò của những thành phần chính này và mối quan hệ giữa chúng.

1.11 Mô hình kiến trúc hệ thống tổng hợp tin tức thể thao

Như đã trình bày ở trên, luận án lựa chọn tiếp cận ứng dụng công nghệ Web ngữ nghĩa để nghiên cứu phát triển hệ thống tổng hợp tin tức thể thao. Ý tưởng cốt lõi là biểu diễn tin tức sử

dụng mô hình có ngữ nghĩa và sau đó xử lý tin tức dựa trên khai thác mô hình ngữ nghĩa đó. Ba nhiệm vụ nghiên cứu cụ thể là (1) phương pháp sinh chú thích ngữ nghĩa từ các tin tức thể thao, (2) phương pháp tìm kiếm ngữ nghĩa tin tức thể thao với câu truy vấn được diễn đạt bởi ngôn ngữ tự nhiên và (3) gợi ý tin tức dựa trên ngữ nghĩa cho hệ thống tin tức thể thao. Kết quả của mỗi nghiên cứu trên là một thành phần quan trọng nằm trong hệ thống cần nghiên cứu. Bên cạnh đó, hệ thống cần những thành phần khác để hoạt động như một hệ thống tin tức hoàn chỉnh. Kiến trúc tổng thể của hệ thống tin tức thể thao có ngữ nghĩa được mô tả ở hình 1.7 như sau:



Hình 1.7 *Kiến trúc tổng thể của hệ thống BKSport*

Hệ thống tổng thể được đặt tên là BKSport. Đối với người dùng, hệ thống hoạt động như trang tin tức thông thường mà ở đó người dùng có thể xem tin tức tổng hợp từ một số nguồn tin cậy, được hỗ trợ tính năng tìm kiếm và gợi ý tin tức.

1.11.1 Crawler

Hệ thống con crawler là thành phần đầu tiên trong kiến trúc của hệ thống, có nhiệm vụ thu thập các tin tức thể thao từ nhiều nguồn trên Web một cách tự động. Kết quả của quá trình thu thập được lưu trữ dưới dạng HTML, tiếp theo nó được đưa sang các thành phần khác để xử lý trong các công đoạn sau, và hiển thị cho người đọc. Hệ thống con này được thiết kế để hoạt động theo nguyên lý của một Web Crawler chung [109], gồm 2 tác vụ chính. Tác vụ thứ nhất là lấy danh sách các địa chỉ. Dữ liệu cơ sở phải có độ chính xác cao, nên từ một hay nhiều địa chỉ Website ban đầu tương ứng với trang chủ của các Website thể thao nổi tiếng bằng tiếng Anh như BBC Sport, SkySports, ESPN, crawler phải lấy được một danh sách địa chỉ các trang Web liên kết trực tiếp hay gián tiếp tới danh sách địa chỉ ban đầu. Do các tin tức là các trang Web liên kết với nhau thông qua URL, bản chất của bài toán này là một phép duyệt đệ quy trên một đồ thị mà mỗi đỉnh là một trang Web và mỗi một cạnh là một liên kết. Tác vụ thứ hai của nó là lần lượt tải về tất cả nội dung cần quan tâm có trong các địa chỉ nằm trong danh sách trên.

Trong các tính năng cần có của Crawler được nêu trong [110], thì tính tươi mới, chất lượng và tính bao phủ của thông tin là quan trọng nhất. Luận án không đặt mục tiêu nghiên cứu kỹ thuật mới trong thu thập tin tức để phát triển web crawler mà sử dụng các kỹ thuật đã được công bố. Các tin tức thể thao được lấy về nhờ crawler sẽ được phân tích và giữ lại những thông tin cần thiết như tiêu đề, mô tả, nội dung, đường dẫn ảnh… Sau đó, chúng được lưu vào cơ sở dữ liệu tin tức (News Database) để thành phần sinh chú thích ngữ nghĩa có thể truy cập và để được hiển thị trong hệ thống cho người dùng cuối.

1.11.2 Ontology thể thao

Trong các hệ thống ứng dụng Web ngữ nghĩa, ontology luôn đóng vai trò thiết yếu. Ontology cung cấp tập từ vựng có kiểm soát, được định nghĩa một cách chặt chẽ để biểu diễn thông tin, dữ liệu, tri thức của miền lĩnh vực ứng dụng đang quan tâm. Trong hệ thống tin tức thể thao đề xuất, ontology thể thao là rất quan trọng, được xây dựng đầu tiên và được đặt tên là BKSport. Ontology này cung cấp tập từ vựng để tạo lên mô hình ngữ nghĩa của hệ thống. Ontology BKSport nằm ở trung tâm của hệ thống và đóng vai trò dẫn dắt sự hoạt động các thành phần quan trọng khác của hệ thống. Cụ thể ở thành phần chú thích ngữ nghĩa, nơi tạo ra chú thích ngữ nghĩa các tin tức thể thao, các thuật toán sinh chú thích ngữ nghĩa phải khai thác các thành phần từ vựng trong ontology BKSport và kết hợp với các nội dung trong văn bản để tạo ra ngữ nghĩa trong quá trình sinh chú thích ngữ nghĩa.

Trong thành phần cổng thông tin, một chức năng quan trọng là tìm kiếm ngữ nghĩa một cách thông minh, chính xác thì ontology BKSport được sử dụng trong diễn đạt câu hỏi ngữ nghĩa để máy tính hiểu được. Trong nghiên cứu chức năng tìm kiếm ngữ nghĩa bằng các câu hỏi tự nhiên của luận án, vấn đề đặt ra là chuyển đổi trong suốt câu hỏi ngôn ngữ tự nhiên sang dạng truy vấn ngữ nghĩa, các phương pháp thuật toán để thực hiện điều này phải làm theo cách nào đó sao cho tương ứng được giữa các thành phần của ngôn ngữ tự nhiên với các thành phần có trong ontology BKSport.

Với vai trò quan trọng của ontology BKSport trong kiến trúc hoạt động hệ thống tổng hợp tin tức thể thao nêu trên, tác giả thấy rằng sự thành công của hệ thống phụ thuộc vào chất lượng của ontology BKSport được xây dựng lên. Chất lượng của ontology BKSport quyết định chất lượng và hiệu quả làm việc của các thành phần khác trong hệ thống. Một ontology BKSport chất lượng cao cần đảm bảo các yêu cầu: đầy đủ, chính xác, không dư thừa, không nhập nhằng.

- Đầy đủ để có thể diễn đạt các thông tin về các thành phần mình cần như các mảnh thông tin, nội dung, sắc thái, ý nghĩa của tin tức thể thao …
- Chính xác để diễn đạt đúng các quan hệ giữa các thành phần của thông tin, nếu không chính xác sẽ dẫn đến sai lệch các thuật toán xử lý, diễn đạt sai các ý nghĩa
- Và không dư thừa, nhập nhằng. Chức năng quan trọng của hệ thống ngữ nghĩa là giúp cho máy tính hiểu đúng yêu cầu của người dùng …muốn vậy các khái niệm không được nhập nhằng, và không dư thừa.

1.11.3 Sinh chú thích ngữ nghĩa

Hệ thống sinh chú thích ngữ nghĩa là một thành phần quan trọng trong cấu trúc hệ thống tổng hợp tin tức thể thao mà luận án cần xây dựng. Nó thể hiện ý tưởng cốt lõi và cũng là nhiệm vụ trọng tâm trong nghiên cứu của luận án. Một cách khái quát, thành phần sinh chú thích ngữ nghĩa chịu trách nhiệm tạo nên các chú thích ngữ nghĩa cho các tin tức thể thao với đầu vào là các tài liệu HTML do crawler lấy về.

Trong luận án này, một chương lớn được dành để trình bầy việc xây dựng, phát triển thành phần này và đồng thời để thành phần này hoạt động được một cơ sở tri thức lĩnh vực thể thao đủ lớn cần được xây dựng. Luận án sẽ nghiên cứu các phương pháp sinh chú thích ngữ nghĩa và cũng quan tâm đến các loại chú thích ngữ nghĩa khác nhau. Ví dụ: chú thích các thực thể nào? chú thích liên quan đến con người thể thao nào? đến sự kiện thể thao nào? về chủ đề thể thao nào? về lời tuyên bố của các nhân vật thể thao nào? về kết quả thi đấu của một trận bóng đá? về thông tin chuyển nhượng của cầu thủ? ...

Tác giả cài đặt các thuật toán sinh chú thích khác nhau, nhưng nó có điểm chung là dựa vào ontology và cơ sở tri thức. Các kỹ thuật phân tích văn bản, xử lý ngôn ngữ tự nhiên được kết hợp với các cấu trúc chú thích ngữ nghĩa khác nhau để tạo ra được các kiểu chú thích khác nhau. Các chú thích ngữ nghĩa được tạo ra dưới dạng metadata n-quad gắn với tin tức thể thao tương ứng, vì vậy nó rời rạc. Chú thích ngữ nghĩa được tạo ra là đầu vào và được lưu giữ trong kho dữ liệu ngữ nghĩa của hệ thống (Annotation Base).

1.11.4 Cổng thông tin ngữ nghĩa

Cổng thông tin là bộ mặt mà người dùng nhìn nhận về hệ thống. Giống như các trang tin thể thao thông dụng khác, nó gồm đầy đủ các tính năng cơ bản giúp cho người dùng xem tin tức, duyệt, liên kết các tin tức liên quan. Nó cung cấp môi trường tích hợp để người dùng có thể truy cập dễ dàng vào các nguồn tin tức thể thao không thuần nhất bằng máy tính để bàn, máy tính xách tay cũng như thiết bị di động.

Trọng tâm của thành phần này thể hiện kết quả nghiên cứu quan trọng thứ hai của luận án, đó là chức năng tìm kiếm ngữ nghĩa trong trang tin dựa trên việc cài đặt thuật toán tìm kiếm ngữ nghĩa với câu hỏi ngôn ngữ tự nhiên (nội dung của nó sẽ được trình bày chi tiết trong chương 3 của luận án). Để làm được điều đó sẽ phải thực hiện các thuật toán tự động chuyển đổi câu hỏi ngôn ngữ tự nhiên sang dạng câu hỏi ngữ nghĩa. Câu truy vấn ở dạng ngôn ngữ tự nhiên trước tiên sẽ được bộ chuyển đổi truy vấn (Query Transformer) chuyển về câu truy vấn SPARQL. Bộ trả lời truy vấn (Query Answering) nhận câu truy vấn SPARQL và truy vấn vào thành phần cơ sở chú thích (Annotation Base) và cơ sở dữ liệu tin tức (News Database) để lấy ra tin tức và các thông tin liên quan phục vụ hiển thị kết quả cho người dùng. Kết quả trả về cho người dùng là các tin tức phù hợp với ý nghĩa (ngữ nghĩa) của câu hỏi.

Ngoài ra, một thành phần quan trọng khác của cổng thông tin ngữ nghĩa là phân hệ gợi ý tin tức (Recommender). Thành phần này có chức năng tự động gợi ý các tin tức khác có liên quan về ngữ nghĩa và nội dung với tin tức mà người dùng đang đọc. Chương 4 sẽ trình bày kết quả nghiên cứu của luận án liên quan đến việc phát triển phân hệ này. Hướng nghiên cứu là đề xuất một thuật toán dựa trên yếu tố ngữ nghĩa để có thể gợi ý không chỉ các tin tức cùng chủ đề với tin đang đọc mà còn có thể gợi ý những tin tức nói về các thực thể có quan hệ ngữ nghĩa với tin tức mục tiêu.

1.11.5 Mô tơ suy diễn và tìm kiếm ngữ nghĩa

Mô tơ suy diễn và tìm kiếm ngữ nghĩa là một loại máy ngữ nghĩa đặc thù. Nó là một phần mềm, sản phẩm của công nghệ thông tin, thường được phát triển bởi các cộng đồng khoa học công nghệ hoặc công ty phần mềm lớn. Trong hệ thống tổng hợp tin tức thể thao của luận án, thành phần này phục vụ một cách tự động cho nhiều thành phần khác như sinh chú thích ngữ nghĩa, tìm kiếm ngữ nghĩa. Nó chịu trách nhiệm về tìm kiếm ngữ nghĩa và các phép xử lý, tính toán suy diễn trên mọi dữ liệu ngữ nghĩa bao gồm chú thích ngữ nghĩa và ontology. BKSport ontology cùng với một tập luật được xây dựng dựa trên cơ sở tri thức về thể thao cũng được nạp vào mô tơ này.

Khi nhận được các câu truy vấn ngữ nghĩa, mô tơ tìm kiếm Allegrograph dựa trên khả năng hiểu được các thuật ngữ, cũng như các ý nghĩa, ngữ cảnh của câu truy vấn, nó thực hiện việc tìm kiếm một cách chính xác trong cơ sở chú thích ngữ nghĩa các tin tức thể thao để lấy ra các tin tức phù hợp với câu truy vấn rồi gửi về hệ thống.

Đã có nhiều nghiên cứu chuyên sâu về việc phát triển mô tơ ngữ nghĩa và đã xuất hiện các sản phẩm thương mại hóa. Hệ thống của luận án sử dụng mô tơ ngữ nghĩa Allegrograph được xây dựng bởi công ty phần mềm Franz để suy diễn và tìm kiếm ngữ nghĩa. Tác giả không đi sâu nghiên cứu về mô tơ suy diễn và tìm kiếm ngữ nghĩa vì đây là một công việc đòi hỏi có sự đầu tư và nghiên cứu trên quy mô rất lớn.

1.11.6 Kho dữ liệu ngữ nghĩa

Kho dữ liệu ngữ nghĩa của hệ thống là bộ chứa lưu trữ và quản lý tất cả các cơ sở chú thích ngữ nghĩa về các tin tức thể thao được sinh ra. Nó cũng là nơi cung cấp dữ liệu cho mô tơ tìm

kiếm ngữ nghĩa. Việc quản lý và bảo trì kho dữ liệu đòi hỏi người quản lý có chuyên môn, theo dõi thường xuyên, bởi vì dữ liệu của nó được cập nhật và bổ sung liên tục.

Trong hệ thống của luận án, kho dữ liệu lưu trữ các dữ liệu dưới dạng bộ ba RDF. Hiện tại, kho dữ liệu của luận án được xây dựng trên nền tảng của Allegrograph Framework. Người dùng muốn khai thác thủ công kho dữ liệu thông qua Web View (giao diện Web) có thể tra cứu, tìm kiếm thông tin trên đó. Tuy nhiên, với cách thức này, kết quả trả về được xử lý hoàn toàn bằng con người. Đối với các nhà phát triển các dịch vụ Web hoặc nhà lập trình, họ khai thác các dữ liệu trong kho này một cách tự động thông qua máy tìm kiếm ngữ nghĩa. Thành phần này còn cho phép khai thác tương tác trực tiếp với nhiều giao diện khác nhau.

1.12 Kết luận chương

Trong chương này luận án đã trình bày một cách tóm tắt các kiến thức nền tảng cho nội dung các chương tiếp theo. Mục 1.1 giới thiệu về nguồn gốc, khái niệm và kiến trúc của Web ngữ nghĩa. Mục 1.2 và mục 1.3 dành sự quan tâm đặc biệt đến ontology, ngôn ngữ biểu diễn ontology và dữ liệu ngữ nghĩa là các kiến thức sẽ được áp dụng cho chương tiếp theo. Mục 1.4 luận án đề cập đến tìm kiếm ngữ nghĩa để tìm ra phương pháp cải thiện độ chính xác của tìm kiếm. Mục 1.5 quan tâm đến kho dữ liệu ngữ nghĩa mở. Mục 1.6, mục 1.7, và 1.8 trình bày về một số lĩnh vực ứng dụng Web ngữ nghĩa, một số nghiên cứu Web ngữ nghĩa tiêu biểu, và website và cổng thông tin tin tức có ngữ nghĩa. Mục 1.9 đề cập đến các ứng dụng Web ngữ nghĩa trong lĩnh vực thể thao. Trong mục 1.10, tác giả đề xuất tiếp cận Web ngữ nghĩa xây dựng hệ thống tin tức thể thao. Mô hình kiến trúc hệ thống tổng hợp tin tức thể thao được trình bày trong mục 1.11. Cuối cùng, mục 1.12 là kết luận chương.

Để thực hiện nhiệm vụ của luận án là xây dựng hệ thống tổng hợp tin tức thể thao ứng dụng Web ngữ nghĩa, tác giả đề xuất xây dựng hệ thống tổng hợp tin tức thể thao với ba trọng tâm nghiên cứu là sinh chú thích ngữ nghĩa, tìm kiếm ngữ nghĩa, và gợi ý tin tức. Trên cơ sở kết quả của ba hướng nghiên cứu này, một cổng thông tin ngữ nghĩa của hệ thống được xây dựng dựa trên cơ sở việc áp dụng cách thành tựu tiên tiến của ba hướng nghiên cứu nêu trên. Chương này là cơ sở để các chương tiếp theo đi vào trình bày các công việc cụ thể và các kết quả nghiên cứu cho những nhiệm vụ mà luận án đặt ra.

CHƯƠNG 2. SINH CHÚ THÍCH NGỮ NGHĨA CHO TIN TỨC THỂ THAO

Chương này trình bày những nghiên cứu về sinh chú thích ngữ nghĩa cho tin tức thể thao, đây là nhiệm vụ nghiên cứu đầu tiên của luận án. Sau trình bày cơ sở lý thuyết của bài toán sinh chú thích ngữ nghĩa cho tài liệu và các nghiên cứu liên quan, luận án đề xuất một phương pháp mới cho phép tạo ra các chú thích về tin tức thể thao với các ngữ nghĩa đặc thù và cần thiết cho hệ thống tổng hợp tin tức. Phương pháp cải tiến hiệu quả của tác vụ nhận dạng thực thể có tên trong miền thể thao, sử dụng ontology và cơ sở tri thức. Trên cơ sở đó, luận án đề xuất các thuật toán sinh chú thích ngữ nghĩa cho các tin tức thể thao (cụ thể là tin tức bóng đá) dựa trên việc sử dụng các luật (mẫu) trích chọn. Một số thực nghiệm được tiến hành cho phép đánh giá những hiệu quả đạt được trong các thử nghiệm ở từng nghiên cứu thành phần.

2.1 Đặt vấn đề

Tìm kiếm thông tin chính xác, nâng cao trải nghiệm duyệt đọc tin, tổ chức tin tức một cách phù hợp và phân loại chúng theo các chủ đề là những mục tiêu mà các nhà phát triển các hệ thống tin tức đang hướng đến. Đó cũng là mục tiêu chung của luận án. Như đã thảo luận ở chương trước, hướng tiếp cận mà luận án lựa chọn hứa hẹn mang lại kết quả khả quan đó là ứng dụng công nghệ Web ngữ nghĩa. Ý tưởng xuyên suốt là xây dựng một mô hình biểu diễn thông tin thống nhất và tường minh để thông tin từ nhiều nguồn khác nhau có thể được diễn đạt theo cách mà máy tính có thể "hiểu" và xử lý hiệu quả.

Trong định nghĩa của Tim Berners-Lee về Web ngữ nghĩa, có một phần đề cập trực tiếp đến siêu dữ liệu, chú thích ngữ nghĩa. Có thể thấy rằng, chú thích ngữ nghĩa là một thành phần không thể thiếu trong mọi hệ thống thông tin và phần mềm dựa trên công nghệ ngữ nghĩa. Một trong những tư tưởng quan trọng trong tiếp cận nghiên cứu của luận án là mô hình hóa các tin tức thể thao bằng chính các chú thích ngữ nghĩa của các tin tức đó. Mô hình biểu diễn thông tin có ngữ nghĩa sẽ giúp cho máy tính hiểu được một số ý nghĩa hoặc ngữ cảnh của tin tức. Do đó, để đạt được mục tiêu nghiên cứu của luận án, cần phải giải quyết được bài toán: làm thế nào tạo ra chú thích ngữ nghĩa cho các tin tức thể thao.

Tạo ra các chú thích ngữ nghĩa cho văn bản hay các tài nguyên Web là một vấn đề nghiên cứu quan trọng trong lĩnh vực Web ngữ nghĩa. Đã có nhiều phương pháp được đề xuất, nhưng nhìn chung có thể phân chia vào ba loại: phương pháp thủ công, bán tự động và tự động. Tuy nhiên, chú thích ngữ nghĩa như định nghĩa về nó, bao hàm mô tả "ngữ nghĩa" mà người tạo ra nó muốn mô tả về chủ thể, do đó có những yêu cầu về nội dung biểu đạt phụ thuộc vào lĩnh vực ứng dụng. Ví dụ, với tin tức về một trận đấu bóng đá, ngữ nghĩa quan trọng thường là kết quả của trận đấu hay cầu thủ ghi bàn. Với các tin tức hậu trường, người đọc sẽ quan tâm và muốn tìm kiếm thông tin về tuyên bố hay thái độ của các nhân vật thể thao.

Các nghiên cứu liên quan, với phạm vi áp dụng là lĩnh vực chung hay một vài lĩnh vực cụ thể khác, mới giải quyết một phần yêu cầu của chú thích ngữ nghĩa cho tin tức thể thao. Do đó, luận án tập trung giải quyết thách thức đang tồn tại, nghiên cứu các phương pháp tạo ra những chú thích có khả năng chứa đựng một số ngữ nghĩa đặc thù, cần thiết và là cơ sở cho việc xây dựng các tính năng tìm kiếm, gợi ý tin tức hiệu quả.

Dựa trên cơ sở các công nghệ Web ngữ nghĩa sẵn có, tác giả thấy rằng có thể mô hình hóa các tin tức thể thao bằng chính các chú thích ngữ nghĩa của các tin tức đó. Mô hình biểu diễn thông tin có ngữ nghĩa sẽ giúp cho máy tính hiểu được một số ý nghĩa hoặc ngữ cảnh của tin tức.

Với mục đích trên, chương này có bố cục như sau: sau mục 2.1 Đặt vấn đề, trong mục 2.2, tác giả trình bày một số khái niệm quan trọng về chú thích ngữ nghĩa cho tài liệu và một số nghiên cứu liên quan. Mục 2.3 trình bày nội dung chính tổng hợp những đặc điểm chung phương pháp sinh chú thích ngữ nghĩa trong các nghiên cứu, cũng như giải thích những đóng góp riêng

trong kết quả của từng nghiên cứu. Mục 2.4 giới thiệu kết quả thu được. Mục 2.5 là kết luận chương và các công việc trong tương lai.

2.2 Chú thích ngữ nghĩa cho tài liệu

Chú thích ngữ nghĩa là một tiền đề cơ bản để thực hiện các xử lý có ngữ nghĩa ví dụ, tìm kiếm ngữ nghĩa. Chú thích ngữ nghĩa có quan hệ với nhiều bối cảnh ứng dụng khác nhau, ví dụ như quản lý tri thức y tế, nông nghiệp, truyền thông, thương mại điện tử. Nhiều hệ thống được thực hiện trên quy mô lớn đã triển khai và sử dụng nó.

2.2.1 Khái niệm

Thuật ngữ *"chú thích"* có thể biểu thị cả quá trình chú thích và kết quả của quá trình đó. Khi chúng ta nói "chú thích", chúng ta ám chỉ đến kết quả. Chú thích là gắn một số dữ liệu vào một số dữ liệu khác. Nó thiết lập nên, trong một bối cảnh nào đó, một quan hệ được định kiểu giữa dữ liệu được chú thích và dữ liệu chú thích.

Theo [111] có thể phân biệt ba loại chú thích:

a) Chú thích phi hình thức

b) Chú thích hình thức: định nghĩa một cách hình thức các thành phần và vì vậy máy có thể hiểu được chúng, và

c) Chú thích dựa trên ontology: định nghĩa hình thức các thành phần và chỉ sử dụng các thuật ngữ ontology mà được mọi người hiểu và chấp nhận.

Trong phạm vi của luận án này, tác giả quan tâm đến chú thích dựa trên ontology và tập trung vào chú thích ngữ nghĩa cho tài liệu.

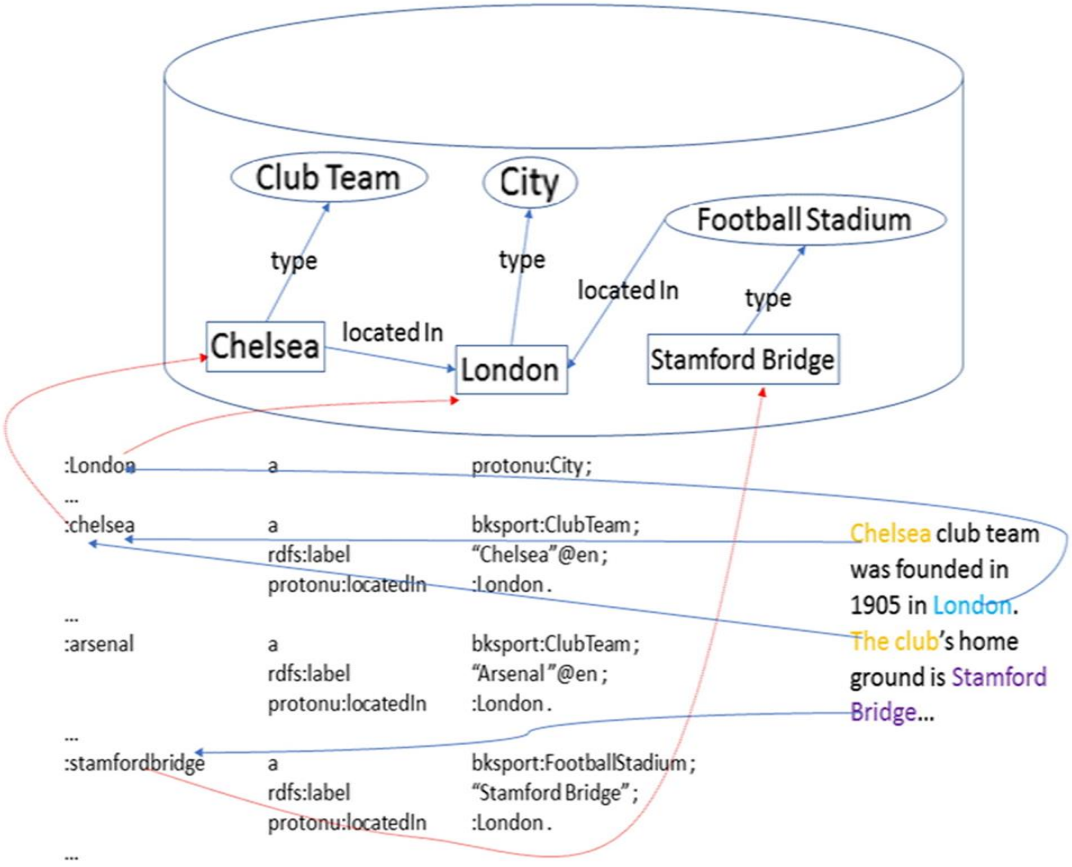
Khi phân tích khái niệm *"chú thích ngữ nghĩa"*. Có nhiều cách hiểu về chú thích ngữ nghĩa tùy theo từng góc độ:

Ở góc độ hành động, chú thích ngữ nghĩa cho tài liệu được hiểu là quy trình tạo ra các mô tả ngữ nghĩa của tài liệu, nó chuyển đổi các cấu trúc cú pháp trong tài liệu thành cấu trúc tri thức. Trong quá trình này, các thực thể trong văn bản được liên kết tới mô tả ngữ nghĩa của chúng thông qua việc kết nối mô hình ngữ nghĩa với ngôn ngữ tự nhiên.

Ở góc độ kết quả, chú thích ngữ nghĩa cho tài liệu sinh ra các siêu dữ liệu cụ thể và lược đồ sử dụng để tạo điều kiện cho tìm kiếm dựa trên khái niệm, suy luận về các tài nguyên Web và trực quan hóa thông tin dựa trên ontology.

Ở góc độ kỹ thuật, chú thích ngữ nghĩa là chú thích về các đề cập đến các khái niệm của ontology (như lớp, thể hiện, thuộc tính, quan hệ) có ở trong văn bản, dựa vào siêu dữ liệu đề cập đến các URI của chúng trong ontology. Cụ thể hơn, chú thích ngữ nghĩa là gắn một thực thể (một chuỗi ký tự, một câu, một đoạn, một phần của một bản ghi hoặc một tài liệu) với một siêu dữ liệu mà ngữ nghĩa của nó được định nghĩa trong một ontology. Chú thích ngữ nghĩa giúp các hệ thống dựa trên Web truyền thống được mở rộng thành các hệ thống dựa Web ngữ nghĩa thông qua việc gắn thêm ngữ nghĩa vào các thông tin có sẵn trong Web truyền thống. Để việc chú thích ngữ nghĩa được phù hợp và chuẩn xác cần đến một ontology. Tập các khái niệm, thuộc tính, quan hệ được định nghĩa trước trong ontology làm cho chú thích ngữ nghĩa trở thành sự diễn đạt một góc nhìn tốt hơn về cấu trúc và nội dung tài liệu, loại bỏ sự nhập nhằng về ngữ nghĩa của tri thức cần mô tả.

Hình 2.1 dưới đây minh họa về chú thích ngữ nghĩa văn bản.



Hình 2.1 *Ví dụ về chú thích ngữ nghĩa*

Chú thích ngữ nghĩa cho tài liệu trên thực tế được hình thức hóa sử dụng các ngôn ngữ RDF hoặc OWL.

2.2.2 Các phương pháp tạo chú thích ngữ nghĩa

Thuật ngữ "chú thích ngữ nghĩa" hiểu theo nghĩa chung nhất là gắn dữ liệu vào một số phần dữ liệu khác. Cho tới hiện tại, chú thích ngữ nghĩa có thể được phân loại là thủ công, bán tự động và tự động hoàn toàn. Nhóm các hệ thống sẽ khác nhau về cấu trúc, phương pháp và công cụ để rút trích thông tin.

Phương pháp chú thích thủ công

Đây là phương pháp đòi hỏi chuyên gia người trực tiếp thực hiện quá trình chú thích các tài nguyên (gắn thực thể với siêu dữ liệu), chuyển các tài nguyên cú pháp hiện có thành các cấu trúc tri thức được liên kết bằng cách thêm thông tin vào văn bản. Một số công cụ phổ biến hỗ trợ chú thích thủ công là CREAM OntoMat, SMORE, Amaya ... Các công cụ biên tập thủ công cho phép con người mô tả ý nghĩa của tài nguyên. Nó đem lại chú thích có chất lượng tin cậy và chính xác hơn so với chú thích tự động.

Tuy nhiên, nhược điểm của phương pháp này là cần nhiều thời gian và công sức, do đó nó thường chỉ được áp dụng trong một số trường hợp ứng dụng đặc biệt như dữ liệu ít hoặc để bổ sung cho phương pháp tự động/bán tự động.

Phương pháp chú thích bán tự động

Đây là phương pháp cần đến sự can thiệp con người ở một mức độ nào đó trong quá trình chú thích để nâng cao chất lượng đầu ra, tuy nhiên một số quá trình đã được tự động hóa. GATE [30] là một công cụ thực hiện chú thích ngữ nghĩa bán tự động. Bộ chú thích NCBO [31]và cTAKE [32] là công cụ khác để thực hiện chú thích ngữ nghĩa bán tự động.

Phương pháp chú thích tự động

Đây là phương pháp không cần đến sự can thiệp của con người trong quá trình chú thích. Phương pháp tự động dựa trên các thuật toán phân tích nội dung tài nguyên để sinh ra các chú thích, và có thể dựa trên các thuật toán thống kê để chú thích ảnh và video. Nó được áp dụng khi cần xử lý dữ liệu ở quy mô lớn. Đây là một nhiệm vụ quan trọng của Web ngữ nghĩa. Siêu dữ liệu Web ngữ nghĩa được tạo ra nhờ các công cụ chú thích tự động với các kết quả tốt nhất dựa trên nhiều thuật toán học máy khác nhau cùng với các tập huấn luyện. Tuy nhiên, các thuật toán này không có khả năng như con người để hiểu được các nội dung có ngữ nghĩa phức tạp, và có thể còn có nhiễu. Vì vậy, các chú thích hiện nay dựa trên các thuật toán tự động cần phải được cải tiến độ chính xác hơn nữa.

Một số công cụ chú thích ngữ nghĩa tự động điển hình là PANKOW [34], C-PANKOW [35], KIM [36]. Trong đó KIM là một nền tảng chú thích tự động dựa trên hệ thống rút trích thông tin GATE [30] với phần mở rộng Annie được nhóm nghiên cứu quan tâm và sử dụng.

2.2.3 Một số nghiên cứu liên quan

Những nghiên cứu đầu tiên tập trung phát triển các hệ thống biên tập chú thích ngữ nghĩa một cách thủ công. Một số ví dụ nổi bật là Semantator [27], M-OntoMat Annotizer [28], Annotea [29], Zemanta (http://www.zemanta.com).

Trong những năm gần đây nhiều nghiên cứu [112] [113] [34] [33] [114] đã được thực hiện để phát triển các hệ thống chú thích ngữ nghĩa tự động và bán tự động. Tuy nhiên, không có hệ thống nào được thiết kế để làm việc cho lĩnh vực thể thao.

Hệ thống Pankow (Pattern-based Annotation through Knowledge on the Web) [34] đã khai thác mô hình bề mặt và sự dư thừa dữ liệu trên Web để tự động phân loại các thực thể trong văn bản sử dụng một ontology có sẵn. Các mô hình là các nhóm từ như <Concept> <Instance> và <Instance> <is_a> <Concept>. Hệ thống xây dựng nên các mô hình này bằng cách nhận dạng tất cả các tên riêng trong văn bản (sử dụng Part-of-Speech Tagger) và kết hợp mỗi tên riêng với một trong 58 khái niệm của ontology vào trong một giả thiết. Sau đó mỗi giả thiết được thử nghiệm với trang Web thông qua các truy vấn Google và số lượng xuất hiện là thước đo để đánh giá độ chính xác của mô hình. Hiệu năng tốt nhất của hệ thống là 24,9% khi hoàn toàn tự động, và 62,09% khi hoạt động dưới sự điều khiển của chuyên gia người.

SemTag [33] là thành phần chú thích ngữ nghĩa của nền tảng Seeker, được dùng để thực hiện việc chú thích các trang Web ở quy mô lớn. Nó làm việc với một ontology hạng nhẹ có tên là TAP, trong đó bao gồm một loạt thông tin từ vựng và phân loại các mục tin thông thường. Sau khi chú thích mọi đề cập có thể của các thực thể từ ontology TAP, SemTag thực hiện thuật toán giải nhập nhằng dựa trên nguyên tắc phân loại. Nó sử dụng một mô hình vectơ không gian để gắn khái niệm đúng hoặc để xác định đề cập này không tương ứng với một khái niệm trong ontology. Độ chính xác tốt nhất của SemTag là khoảng 82%, trong khi đó độ bao phủ chưa được công bố.

Trong [115], các tác giả đã mô tả hệ thống Asknet, một hệ thống trích rút thông tin dành cho việc xây dựng dữ liệu Web ngữ nghĩa quy mô lớn từ văn bản phi cấu trúc. Trình tự trích rút thông tin của Asknet là như sau. Đầu tiên cú pháp của các câu trong văn bản được phân tích bởi bộ phân tích cú pháp C&C. Giai đoạn nhận dạng thực thể có tên được thực hiện bằng cách sử dụng bộ đánh dấu NER C&C. Sau đó các câu được phân tích, Asknet sử dụng một mô tơ phân tích ngữ nghĩa có tên là Boxer để sinh ra các biểu diễn logic bậc một. Hệ thống đạt được độ chính xác tổng thể là 79,1%.

Nghiên cứu của [38] đã đề xuất một thuật toán dựa trên cây hạt nhân để trích rút các quan hệ giữa hai thực thể. Họ đã đề xuất một cây hạt nhân mới, được gọi là "hạt nhân cây được làm giàu chức năng", để vượt qua các vấn đề nhập nhằng trong cây cú pháp truyền thống nhằm nắm bắt quan hệ ngữ nghĩa tốt hơn.

Nhóm tác giả [39] đã giới thiệu một tiếp cận để trích rút các quan hệ giữa các thực thể trong lĩnh vực y học có sử dụng mô hình ngôn ngữ. Họ đã sử dụng MetaMap để trích rút các thực thể có tên trong lĩnh vực y học như tên thuốc, tên bệnh nhân ... Để trích rút các quan hệ mong muốn,

họ đã thiết kế một mô hình ngôn ngữ dựa trên sự lựa chọn các bài báo của PubMed Central. Các thử nghiệm của họ đạt độ chính xác 74,21%.

Nhóm nghiên cứu [114] đã đề xuất một tiếp cận để trích rút các quan hệ ngữ nghĩa giữa các nhóm từ danh từ (các danh định) dựa trên sự phối hợp các thông tin ngữ nghĩa được cung cấp bởi ResearchCyc để xử lý các bộ phân tích cú pháp sơ yếu. Phương pháp đã đạt giá trị đo tổng thể F1 là 77,62% tại SemEval 2010.

Trong mọi hệ thống ứng dụng công nghệ Web ngữ nghĩa, nội dung của chú thích sẽ quyết định các chức năng xử lý thông tin thông minh mà hệ thống cung cấp tới người dùng. Trong luận án này, thông tin ngữ nghĩa trong các chú thích cần hướng đến việc bổ sung "ý nghĩa" về các dữ liệu mà người dùng quan tâm khi tìm kiếm – tra cứu tin tức. Nói cách khác, các chú thích ngữ nghĩa nếu được sinh ra cần biểu đạt được những gì mà các chức năng tìm kiếm ngữ nghĩa hay gợi ý tin tức yêu cầu. Khi truy cập một trang tin thể thao, người đọc thường có ưu tiên muốn tìm kiếm thông tin về kết quả của các sự kiện thể thao như trận đấu, các hành động – hoạt động diễn ra. Họ cũng quan tâm đến các thông tin gắn với các nhân vật, tổ chức thể thao nổi tiếng, các hoạt động chuyển nhượng … Để hệ thống có thể trả lời các câu hỏi như "Đội bóng nào đã đánh bại Barcelona tuần qua?" "Cầu thủ nào đã ghi bàn?" "Chuyện gì diễn ra giữa Ronaldo và Messi?", cần có các chú thích ngữ nghĩa chứa đựng các thông tin tương ứng.

Trong khi đó, kết quả của các nghiên cứu liên quan nói trên chưa đáp ứng được yêu cầu này một cách thỏa đáng. Đầu tiên, trong hệ thống tổng hợp tin tức thể thao tin tức được thu thập từ nhiều nguồn nên có số lượng lớn và có tần suất cập nhật cao. Do đó, phương pháp tạo chú thích thủ công [27] [29] chỉ dành cho biên tập viên với mục đích thẩm định, nâng cao chất lượng của chú thích. Giải pháp này không phù hợp để áp dụng trên tập toàn bộ các tin tức.

Các nghiên cứu [35] [36] cho phép phát hiện các thực thể có tên, nhưng do thiết kế cho bài toán tổng quát nên các phương pháp này chỉ gán các thực thể trên vào các lớp thông tin cơ bản là: Người, Tổ chức, Địa chỉ, Tiền tệ, Thời gian … Trong khi đó, SemTag chỉ sử dụng TAP ontology và không hỗ trợ sử dụng ontology lĩnh vực khác. Các nghiên cứu khác thực hiện tác vụ này trong lĩnh vực đặc thù như y tế, sinh học.

Một số phương pháp hướng đến việc phát hiện quan hệ [114] [38] [39], tuy nhiên vẫn chưa cho phép tạo ra các bộ ba ngữ nghĩa dưới dạng RDF, OWL. Ví dụ, [115] tạo ra các biểu diễn logic bậc một. Phương pháp [39] được xây dựng để áp dụng cho lĩnh vực y học, nó đòi hỏi tri thức miền từ MetaMap và PubMed, do đó không khả thi để áp dụng vào lĩnh vực thể thao.

Với những phân tích đã nêu, tác giả thấy rằng bài toán sinh chú thích ngữ nghĩa trong lĩnh vực đặc thù như thể thao vẫn là một bài toán mở, chưa có lời giải thỏa đáng. Nghiên cứu một phương pháp tự động tạo chú thích ngữ nghĩa cho số lượng lớn tin tức thể thao với thời gian xử lý ngắn và độ chính xác tương đối có ý nghĩa quan trọng.

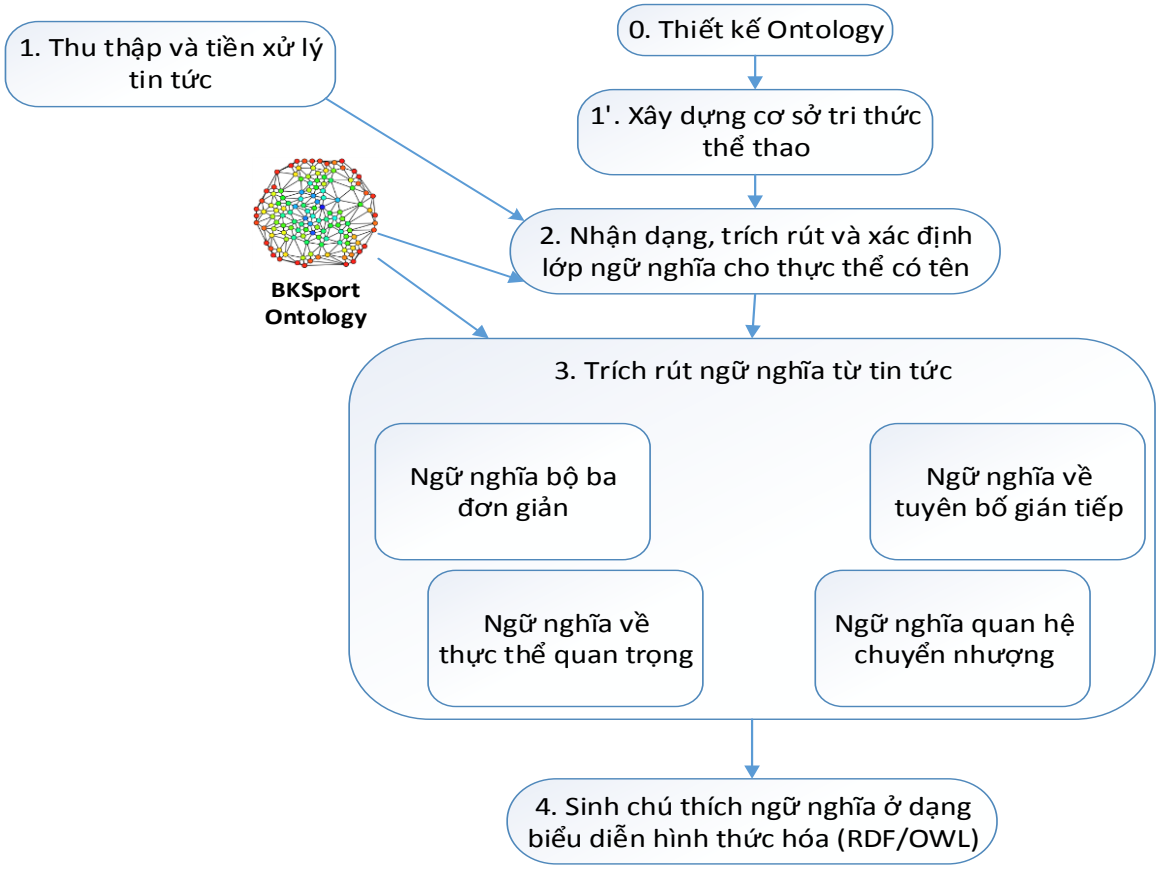
2.3 Một phương pháp sinh chú thích ngữ nghĩa cho tin tức thể thao dựa trên ontology và luật trích chọn

2.3.1 Tổng quan về phương pháp đề xuất

Từ những vấn đề còn tồn tại được nêu ở trên, luận án đề xuất một phương pháp sinh chú thích ngữ nghĩa cho các tin tức thể thao dựa trên việc sử dụng ontology, cơ sở tri thức và luật trích chọn. Để tạo ra chú thích ngữ nghĩa tự động, luận án tập trung vào việc nhận diện các thực thể có tên trong lĩnh vực thể thao. Nếu sử dụng các kết quả nghiên cứu của bài toán nhận dạng thực thể có tên và gắn nhãn, các thực thể này chỉ được xác định như là thể hiện của các lớp chung như Người, Địa điểm, Tổ Chức, Thời gian … và chúng sẽ không có nhiều ý nghĩa trong việc tạo ra ngữ nghĩa hữu ích. Luận án tiến hành nhận dạng chúng trên cơ sở so khớp chúng với các thể hiện trong cơ sở tri thức thể thao, từ đó xác định lớp của chúng là các khái niệm thuộc ontology BKSport được luận án xây dựng. Dựa trên các thể hiện của các thực thể thể thao được phát hiện, một số thuật toán đã được luận án đề xuất để phát hiện các dạng ngữ nghĩa khác nhau. Dựa trên luật trích chọn để xây dựng nên phần lớn các thuật toán là tư tưởng chung của luận án. Những luật này diễn tả mô hình biểu diễn ngữ nghĩa cần trích rút dưới dạng liên kết giữa các thực thể và các quan hệ trong ontology. Có thể nói, mặc dù trình bày trong tổng thể

một phương pháp, nhưng đây là kết quả tổng hợp của nhiều nghiên cứu trải dài trong quá trình thực hiện luận án.

Các giai đoạn trong phương pháp tổng thể được minh họa ở hình 2.2 dưới đây:



Hình 2.2 *Quá trình chú thích ngữ nghĩa*

Phương pháp được chia làm 5 bước, mỗi bước cụ thể được giải thích trong các tiểu mục từ 2.4.2 đến 2.4.6 không kể giai đoạn thu thập tự động tin tức từ nhiều nguồn trên World Wide Web và lưu trữ trong cơ sở dữ liệu.

- Bước đầu tiên là thiết kế và xây dựng một ontology miền ứng dụng mà luận án đề cập tới.
- Xây dựng một cơ sở tri thức về thể thao dựa trên các từ vựng trong ontology.
- Nhận dạng các thực thể có tên, xác định lớp ngữ nghĩa cho các thực thể này. Đối với bước này, luận án đã đề xuất một phương pháp cho phép nhận dạng các thực thể có tên thuộc lĩnh vực thể thao có hiệu quả cao hơn các nghiên cứu liên quan.
- Phát hiện – trích rút ngữ nghĩa từ tin tức thể thao. Thực chất bước này bao gồm một số phương pháp cụ thể được luận án đề xuất nhằm sinh ra những ngữ nghĩa khác nhau trong tin tức thể thao. Các phương pháp này được xây dựng trong quá trình thực hiện luận án, và lần lượt công bố qua các công trình khác nhau. Những ngữ nghĩa mà luận án phát hiện khác biệt với các phương pháp sinh chú thích ngữ nghĩa đề cập trong các nghiên cứu liên quan. Những ngữ nghĩa mới được luận án đưa ra cụ thể là:
  - Ngữ nghĩa bộ ba đơn giản – diễn tả các hoạt động, sự kiện diễn ra trong tin tức.
  - Ngữ nghĩa về tuyên bố gián tiếp.
  - Ngữ nghĩa về chủ đề quan trọng mà tin tức đề cập.
  - Ngữ nghĩa về các hoạt động chuyển nhượng trong tin tức thể thao.

Các tiểu mục dưới đây sẽ trình bày cụ thể nội dung chi tiết của các bước trong phương pháp sinh chú thích ngữ nghĩa cho tin tức thể thao mà luận án đề xuất.

2.3.2 Xây dựng Ontology cho hệ thống

Đầu tiên, có thể khẳng định – việc xây dựng một ontology định nghĩa một cách tường minh và hình thức các thành tố từ vựng đóng vai trò làm nền tảng biểu diễn tri thức trong miền ứng dụng thể thao là một nội dung quan trọng và liên quan tới tất cả các nghiên cứu của luận án. Ontology liên quan đến việc tạo ra chú thích ngữ nghĩa lẫn sinh ra các truy vấn tìm kiếm ngữ nghĩa, ảnh hưởng tới thuật toán gợi ý tin tức. Vì vậy, xây dựng ontology thể thao không phải là một tác vụ chỉ nằm trong quy trình sinh chú thích ngữ nghĩa. Tuy nhiên, nội dung và cách ontology thể thao BKSport được xây dựng có ảnh hưởng lớn tới kết quả của các thuật toán sinh chú thích ngữ nghĩa mà luận án đề xuất. Đó là lý do nội dung này được tác giả quyết định trình bày trong chương 2.

Năm 1993 [62] Gruber đã định nghĩa rằng "ontology là một đặc tả rõ ràng của một khái niệm hóa (được chia sẻ)". Các nguyên tắc cơ bản được định nghĩa bởi Gruber để thiết kế và xây dựng ontology là như sau:

- Rõ ràng và khách quan: các thuật ngữ cần được định nghĩa bằng ngôn ngữ tự nhiên sử dụng ontology một cách rõ ràng và khách quan.
- Tính toàn vẹn: định nghĩa phải đầy đủ và biểu thị ý nghĩa của một thuật ngữ cụ thể.
- Tính nhất quán: không có mâu thuẫn giữa các kết luận phát sinh từ các tri thức lý luận và các ngữ nghĩa của thuật ngữ.
- Tối đa khả năng mở rộng một chiều: không cần thiết phải sửa đổi các thuật ngữ hiện hành khi chúng ta thêm các thuật ngữ khái quát hoặc cụ thể vào trong ontology.
- Tối thiểu các ràng buộc: các ràng buộc trong mô hình nên được giới hạn càng ít càng tốt.

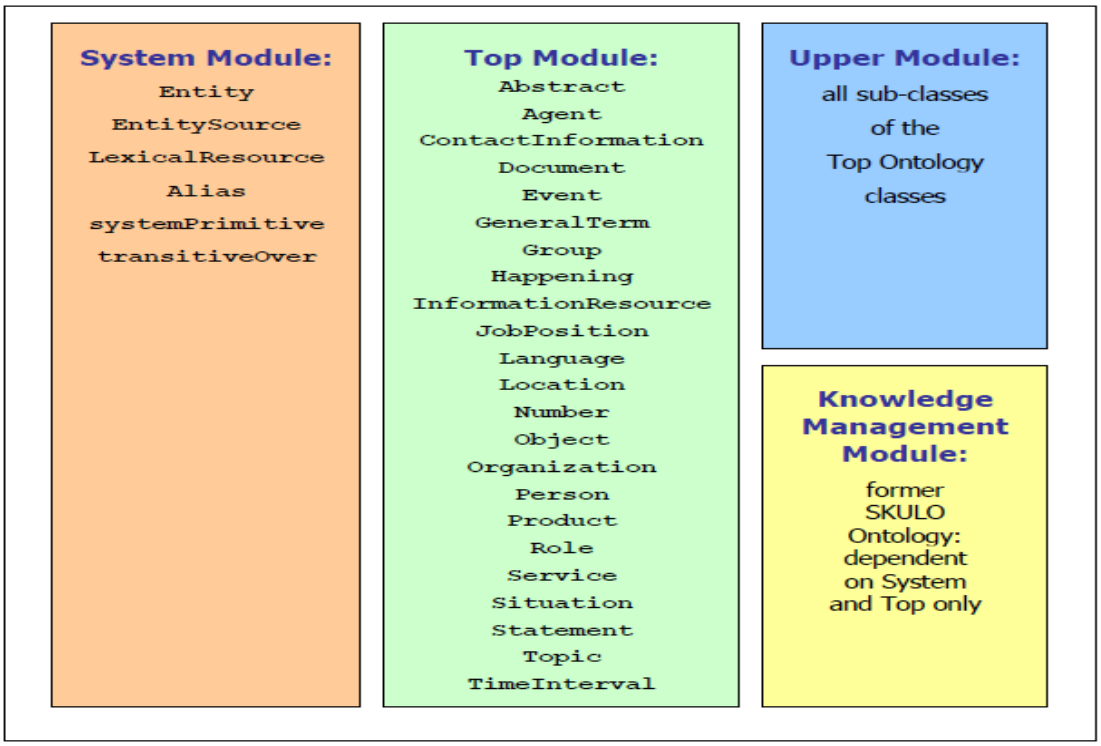
Ontology BKSport được xây dựng trong luận án tuân thủ các nguyên tắc của Gruber. Đồng thời, để mở rộng nền tảng KIM và thực hiện nhận dạng các thực thể có tên trong lĩnh vực thể thao, ontology này cũng được thiết kế để tương thích với ontology PROTON.

2.3.2.1 Ontology PROTON

Ontology PROTON của nền tảng KIM được cải tiến từ ontology KIMO. PROTON được mã hóa bằng ngôn ngữ OWL Lite mạnh mẽ và tiên tiến hơn ngôn ngữ RDFS của KIMO. Nền tảng KIM sử dụng ontology PROTON để chú thích ngữ nghĩa và tìm kiếm đa mô hình cho các tài liệu, dữ liệu, và tri thức. Tổng quan toàn diện về nền tảng KIM được trình bày trong [36].

PROTON chứa khoảng 250 lớp và 100 thuộc tính, cung cấp các khái niệm khái quát cần thiết cho một loạt các tác vụ, bao gồm chú thích ngữ nghĩa, lập chỉ mục và truy hồi tài liệu. PROTON là một ontology có những ưu điểm nổi bật sau: độc lập miền, bao phủ tốt các thực thể có tên về con người, tổ chức, địa điểm, con số, địa chỉ, ngày tháng năm (cơ sở tri thức của nó có khoảng 200.000 mô tả thực thể).

PROTON được tổ chức theo ba cấp với bốn mô-đun độc lập như ở hình 2.3 dưới đây. Mô-đun ontology System chứa các khái niệm cơ bản và trừu tượng nhất. Sau đó, những ontology Top, Upper, và KM (knowledge management) được nâng cấp dựa trên nó để tạo ra kiến trúc mô-đun đặc biệt và đặc trưng của PROTON.

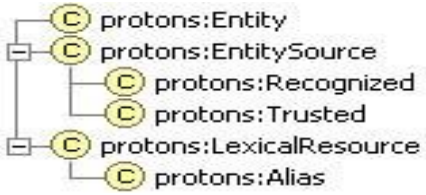


Hình 2.3 *Các mô đun của ontology PROTON*

a) Mô-đun ontology PROTON System (protons.owl)

Ontology PROTON System có sẵn tại http://proton.semanticweb.org/2005/04/protons#. Mô-đun System của PROTON chứa đựng một phân loại của một số siêu từ vựng gốc mà đã được chấp nhận trong một số công cụ đặc thù có chứa thành phần PROTON như công cụ chú thích ngữ nghĩa, công cụ truy cập tri thức. Nó là thành phần duy nhất trong PROTON mà không được thay đổi vì những mục đích mở rộng ontology. Mô đun System có 6 lớp và 7 thuộc tính, được tham chiếu bởi tiền tố "protons:".

6 lớp của mô đun PROTON System là protons:Entity, protons:EntitySource, protons:Recognized, protons:Trusted, protons:LexicalResource, và protons:Alias được mô tả trong hình 2.4 dưới đây. Lớp protons:Entity là gốc "thực sự" của ontology, nó là lớp cha của mô đun PROTON Top với nhiều lớp thực thể đa dạng. Các lớp còn lại của PROTON System có thể xem như là các lớp phụ trợ. Các thể hiện của lớp cha protons:EntitySource được dùng để lấy ra những thông tin tin cậy trong cơ sở tri thức từ những thông tin được trích rút tự động. Lớp protons:Recognized được dùng để xác định một nguồn (chẳng hạn như một chương trình hoặc một mô đun) mà có khả năng nhận dạng và sinh các thực thể từ văn bản với tư cách là bộ phận của tác vụ khai phá dữ liệu hoặc trích rút thông tin. Lớp protons:Trusted được dùng để chỉ ra các thực thể được nhập từ các nguồn "đáng tin cậy" như World Fact Book, các gazetteer của GATE/MUSE/KIM … Lớp protons:LexicalResource được dành riêng cho việc mã hóa các dạng dữ liệu khác nhau như các hậu tố của công ty (ví dụ, "AG", "Ltd."), các tên người (ví dụ, Nicolas Sarkozy, Massaki Shirakawa) mà có liên quan đến quá trình trích rút thông tin và khai phá dữ liệu. Lớp protons:Alias là một lớp quan trọng, nó được dùng để diễn tả các tên của các thể hiện của lớp protons:Entity.



Hình 2.4 *Hệ thống phân lớp của mô đun PROTON System*

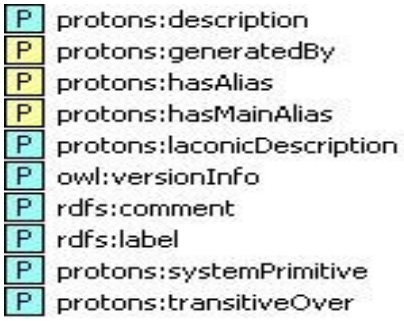
7 thuộc tính của mô đun PROTON System là protons:description, protons:laconicDescription, protons:generatedBy, protons:hasAlias, protons:hasMainAlias, protons:systemPrimitive, protons:transitiveOver.

Hai thuộc tính chú thích là protons:systemPrimitive và protons:transitiveOver. Thuộc tính protons:systemPrimitive được dùng để mã hóa các thông tin của hệ thống cùng với các thể hiện của chúng và các thông tin có liên quan. Những thông tin này không được trình bày cho người dùng cuối, tuy nhiên các mô đun giao diện người dùng và trực quan hóa trong thực tế có thể lọc ra các từ gốc như vậy. Thuộc tính protons:transitiveOver cho biết một thuộc tính là bắc cầu đối với một thuộc tính khác, do đó nhờ nó ta có thể thực hiện được việc mô hình hóa một mẫu hình đặc thù. Ngữ nghĩa của nó được định nghĩa dựa vào tiên đề sau: (p, transitiveOver, q) (x, p, y) (y, q, z) ⇒ (x, p, z). Ví dụ về cách sử dụng protons:transitiveOver như sau: (locatedIn, transitiveOver, subRegionOf) (OldTraffordStadium, locatedIn, Manchester) (Manchester, subRegionOf, England) ⇒ (OldTrafford, locatedIn, England)

Hai thuộc tính dữ liệu là protons:description và protons:laconicDescription. Thuộc tính protons:description được dùng để trình bày mô tả văn bản của một thực thể ở dạng văn bản phi cấu trúc diễn tả bằng ngôn ngữ tự nhiên. Thuộc tính protons:laconicDescription được dùng để trình bày một mô tả ngắn gọn (thường chỉ là 1 câu) về một thực thể. protons:laconicDescription là thuộc tính con của protons:description.

Ba thuộc tính đối tượng là protons:generatedBy, protons:hasAlias, và protons:hasMainAlias. Thuộc tính protons:generatedBy được dùng để xác định bên mà đã đưa thực thể vào trong cơ sở tri thức tương ứng, nó liên kết siêu lớp protons:Entity với siêu lớp protons:EntitySource của mô đun PROTON System. Thuộc tính protons:has Alias được dùng để đề cập đến bí danh của thực thể, nó liên kết lớp protons:Entity với lớp protons:Alias. Thuộc tính protons:hasMainAlias được dùng để đề cập đến bí danh chính thức (tức là bí danh quan trọng nhất) của một thức thể, nó là thuộc tính con của protons:has Alias.

Các thuộc tính của mô đun PROTON System được minh họa trong hình 2.5 dưới đây.



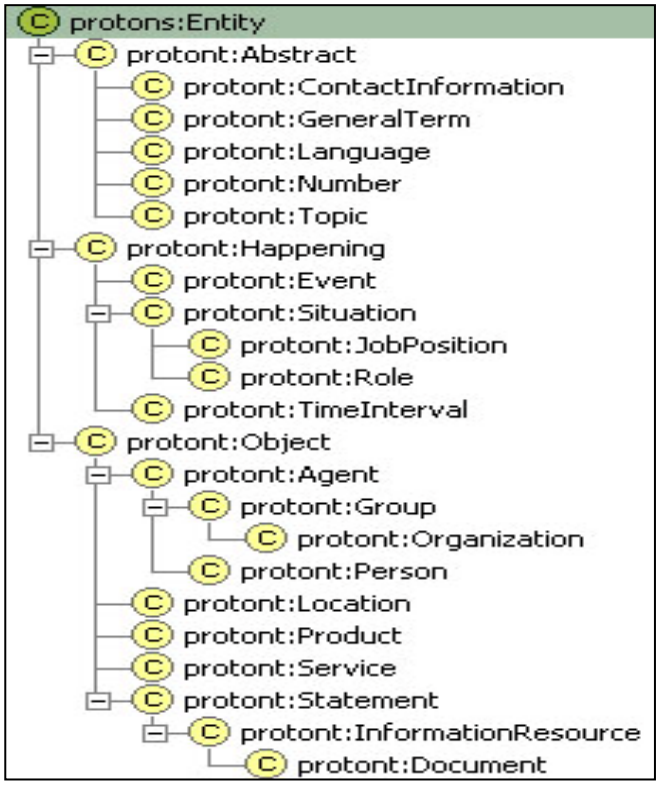
Hình 2.5 *Các thuộc tính của mô đun PROTON System*

b) Mô đun ontology PROTON Top (protont.owl)

Ontology PROTON Top có sẵn tại http://proton.semanticweb.org/2005/04/protont#.

Mô-đun Top của PROTON là mức khái niệm cao nhất, tổng quát nhất, bao gồm trên 20 lớp. Mô đun ontology PROTON Top được tham chiếu bởi tiền tố "protont:" bắt đầu với 3 lớp thiết yếu nhất: protont:Object, protont:Happening, và proton:Abstract (chúng đều là lớp con của lớp cha protons:Entity). Lớp protont:Object chứa các thực thể hiện hữu như tác nhân, địa điểm, phương tiện giao thông. Lớp protont:Happening chứa các thực thể của sự kiện và tình huống. Lớp proton:Abstract chứa các thực thể trừu tượng mà không là đối tượng (Object) cũng như biến cố (Happening). Sau đó, ba lớp nêu trên được chuyên biệt hóa hơn nữa để có các lớp con mô tả về các loại thực thể thế giới thực trọng yếu có tầm quan trọng chung như: cuộc họp, sự kiện thể thao, vị trí việc làm, chính phủ, tổ chức, con người, địa điểm, con số, thời gian, tiền bạc, các giá trị cụ thể… Ngoài ra, các loại thực thể đó cũng có những thuộc tính và quan hệ đặc trưng được mô đun PROTON Top hỗ trợ như protont:subRegionOf, protont:hasPosition, protont:locatedIn, protont:hasMember …

Hình 2.6 dưới đây giúp ta nắm bắt tóm lược về mô đun ontology PROTON Top.



Hình 2.6 *Tóm lược mô đun ontology PROTON Top*

c) Mô đun ontology PROTON Upper (protonu.owl)

PROTON Upper có sẵn tại http://proton.semanticweb.org/2005/04/protonu#.

Mô đun PROTON Upper được tham chiếu bởi tiền tố "protonu:" nằm ở lớp thứ ba của ontology PROTON. Nó là sự mở rộng của mô đun PROTON Top. Các lớp, thuộc tính, và tiên đề của PROTON Upper là các nhánh con của các thành phần tương ứng trong mô đun PROTON Top. Ví dụ, lớp protonu:Mountain là lớp con của lớp protont:Location, lớp protonu:ResourceColection là lớp con của lớp protont:InformationResource.

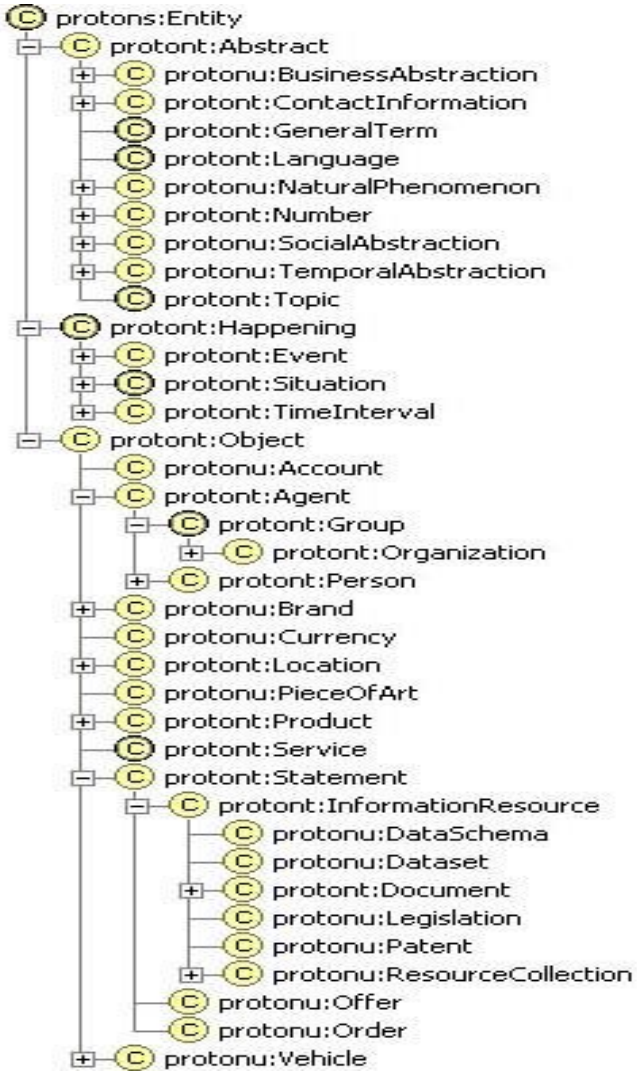
Mô đun PROTON Upper bao phủ hơn 200 lớp thực thể tổng quát mà thường xuất hiện trong nhiều lĩnh vực phổ biến như các loại tổ chức khác nhau, hàng loạt các địa điểm v.v.

Một số lớp của ontology PROTON upper là protonu:BusinessAbstraction, protonu:Address, protonu:NaturalPhenomenon, protonu:SocialAbstraction, protonu:TemporalAbstraction, protonu:Meeting, protonu:JobTitle, protonu:Sport, protonu:Chairman, protonu:President, protonu:SportEvent, protonu:OlympicGames, protonu:SportGames, protonu:Tournament, protonu:Date, protonu:SportOrganization, protonu:SportClub, protonu:SoccerClub, protonu:Team, protonu:Man, protonu:Woman, protonu:Building, protonu:SportBuilding, protonu:Stadium, protonu:Country …

Một số thuộc tính đối tượng của PROTON Upper là protonu:hasCapital, protonu:hasProfession, protonu:hasTitle, protonu:officialPositionIn …

Một số thuộc tính dữ liệu của PROTON Upper là protonu:datePublished, protonu:hasUnit, protonu:ISBN, protonu:ISSN, protonu:stockExchangeIndex …

Hình 2.7 dưới đây cho ta cái nhìn tóm lược mô đun ontology PROTON Upper.



Hình 2.7 *Tóm lược mô đun ontology PROTON Upper*

d) Mô đun ontology PROTON KM (protonkm.owl)

PROTON KM có sẵn tại http://proton.semanticweb.org/2005/04/protonkm#.

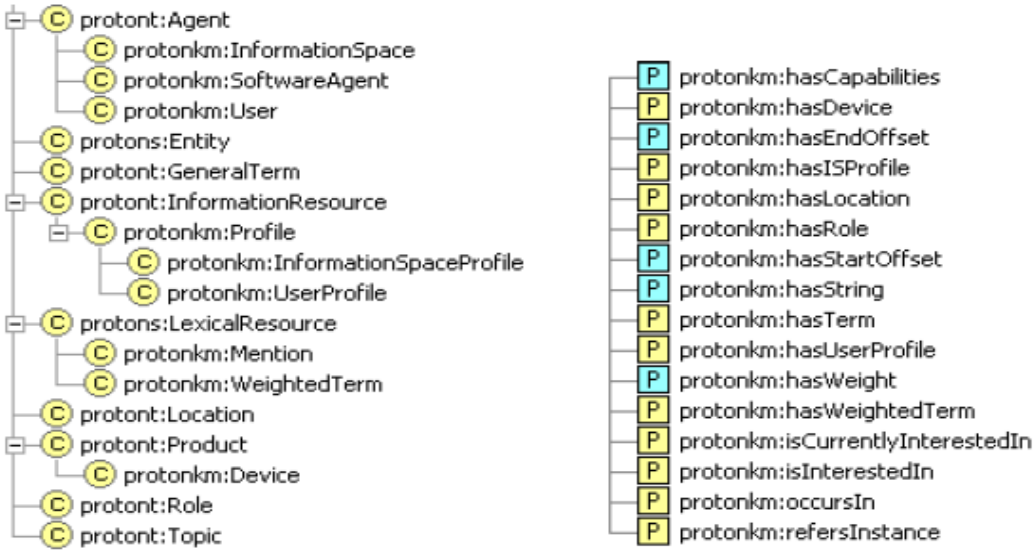
Mô đun PROTON KM (Knowledge Management) được phát triển từ ontology SKULO chứa 38 lớp thực thể chuyên dụng dành riêng cho các tác vụ và ứng dụng quản lý tri thức điển hình. PROTON KM được tham chiếu thông qua tiền tố "protonu:". Mô đun PROTON KM Module phụ thuộc vào mô đun PROTON Sytem và PROTON Top.

Một số lớp của mô đun PROTON KM là protonkm:InformationSpace, protonkm:SoftwareAgent, protonkm:Profile, protonkm:InformationSpaceProfile, protonkm:User, protonkm:UserProfile, protonkm:Mention …

Một số thuộc tính dữ liệu của mô đun PROTON KM là protonkm:hasSartOffset, protonkm:hasEndOffset, protonkm:hasString …

Một số thuộc tính đối tượng của mô đun PROTON KM là protonkm:occursIn, protonkm:refersInstance …

Hình 2.8 mô tả các lớp và thuộc tính của mô đun PROTON KM.



Hình 2.8 *Các lớp và thuộc tính của mô đun PROTON KM*

2.3.2.2 Ontology thể thao của hãng BBC

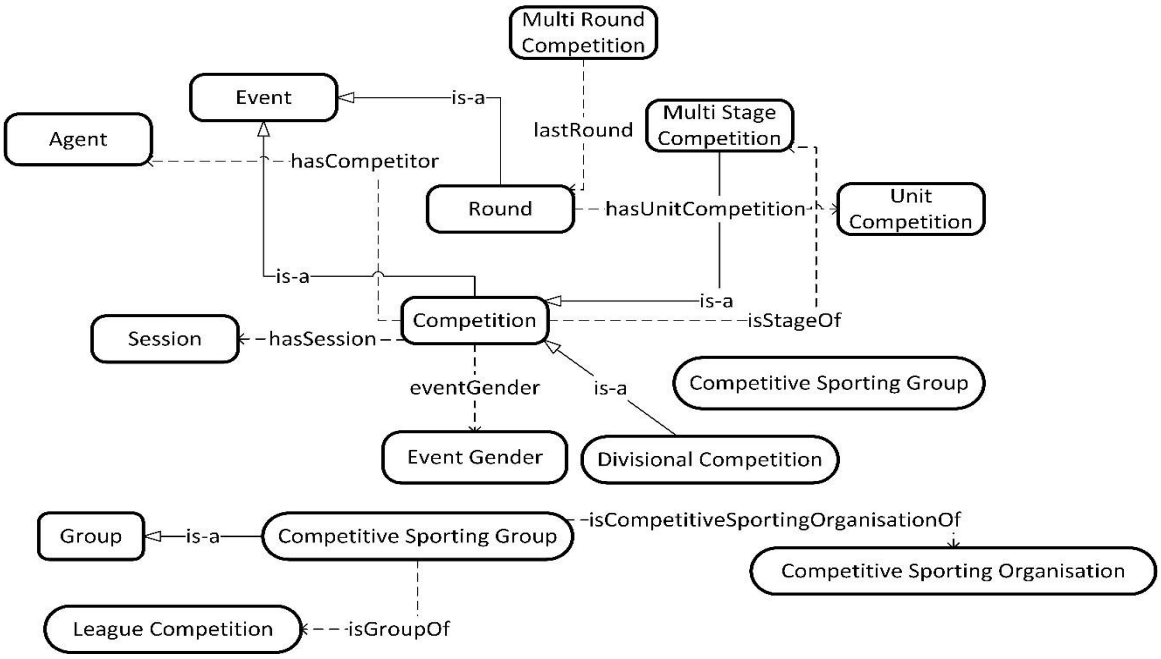
Ontology BBC Sport có sẵn tại http://www.bbc.co.uk/ontologies/sport/2011-02-17.shtml.

BBC là hãng truyền thông dịch vụ công đầu tiên đã xây dựng Website Giải vô địch bóng đá thế giới FIFA World Cup 2010 theo kiến trúc xuất bản ngữ nghĩa động [104]. Hãng truyền thông BBC [106] [107] đã có những nghiên cứu đầu tiên về sử dụng ontology.

Ontology thể thao của hãng BBC là một ontology hạng nhẹ và đơn giản được dùng để xuất bản dữ liệu về các sự kiện và các cuộc thi đấu thể thao. Nó mô tả nhiều khía cạnh đa dạng của các cuộc thi đấu thể thao như giải thưởng của một cuộc thi đấu và làm thế nào để nhận được nó, cạnh tranh giữa các tác nhân trong một cuộc thi đấu, các quy tắc của một cuộc thi đấu thể thao, cấu trúc của một giải đấu thể thao …

Trong mô hình đồ thị của ontology BBC ngay dưới đây, tác giả xác định một số lớp (khái niệm) và thuộc tính là hữu ích và cần thiết nhất để mô tả tri thức ngữ nghĩa trong tin tức thể thao để từ đó chúng được tái sử dụng và kế thừa vào ontology BKSport.

Hình 2.9 mô tả một phần của ontology thể thao của hãng BBC nước Anh.



Hình 2.9 *Một phần của ontology thể thao của hãng BBC*

2.3.2.3 Xây dựng Ontology BKSport

Để phục vụ các tác vụ sau này như nhận dạng thực thể có tên trong tin tức, phát hiện ngữ nghĩa và tìm kiếm ngữ nghĩa, luận án xây dựng một ontology dành cho lĩnh vực thể thao được đặt tên là BKSport.

Ontology BKSport được thiết kế sao cho thỏa mãn các yêu cầu sau: đầy đủ từ vựng để mô tả các thông tin cơ bản trong lĩnh vực thể thao, khả năng cung cấp các khái niệm và các thuộc tính để biểu diễn các sự kiện quan trọng trong các tin tức thể thao như Gruber đã nêu trên. Một ví dụ là "Rashford is a forward, and plays for Manchester United, head coach of which is Jose Mourinho". Một ví dụ khác là "Real Madrid defeats Barcelona with score 2-0".

Ngoài ra, ontology BKSport cũng phải tương thích với ontology PROTON để có thể tái sử dụng nền tảng trích rút thông tin KIM [36]. KIM được xây dựng để phục vụ trong miền mở, nó được trang bị một ontology ở mức cao (ontology PROTON) và một cơ sở tri thức chứa một số lượng lớn các thực thể có tầm quan trọng chung. Được phát triển trong khuôn khổ của dự án SEKT (www.sektproject.com/), PROTON định nghĩa khoảng 250 khái niệm và 100 thuộc tính có thể cung cấp gần như tất cả những khái niệm cần thiết ở mức cao cho chú thích, lập chỉ mục và tìm kiếm ngữ nghĩa. Tuy nhiên, ontology PROTON chỉ định nghĩa các khái niệm và các thuộc tính tổng quát. Do đó, PROTON được lựa chọn là nền tảng để xây dựng ontology BKSport của luận án. Ontology BKSport được dùng để nhận dạng tự động thực thể và trích rút thông tin từ văn bản, dùng để chú thích ngữ nghĩa ở mức cụ thể hơn. Ontology PROTON được mở rộng để có thể chứa đựng các tri thức khái niệm được mã hóa trong các tập dữ liệu phổ biến nhất của dữ liệu mở liên kết như DBpedia, GeoNames v.v. Bốn nhóm từ vựng quan trọng được thừa kế để mô hình hóa ngữ nghĩa các tin tức thể thao là Person, Organization, Location và Time. Vì vậy, ontology BKSport có các lớp và thuộc tính ở mức thấp và chi tiết hơn được tích hợp vào ontology PROTON của KIM.

Ontology BKSport giữ vai trò quyết định trong hệ thống chú thích ngữ nghĩa, nó mô tả các thực thể trong thế giới thực của các môi trường và lĩnh vực thể thao cũng như các đặc tính và các mối quan hệ giữa chúng. Nó cần có một tập từ vựng đầy đủ để mô tả các thông tin cơ bản trong các tin tức thể thao và tin tức chuyển nhượng. Để đạt được điều đó, bên cạnh việc dựa trên PROTON và kết quả khảo sát một số ontology thể thao, luận án cố gắng tái sử dụng một số thuật ngữ phù hợp từ ontology BBC. Các khái niệm và thuộc tính được tái sử dụng này có mức độ chi tiết ngữ nghĩa ở mức độ trung bình, nằm giữa mức khái quát của tập từ vựng kế thừa từ PROTON nhưng chưa đủ chi tiết để diễn tả hết các ngữ nghĩa trong tin tức thể thao.

Cuối cùng, thực hiện tác vụ phân tích các tin tức bóng đá, luận án định nghĩa các khái niệm quan trọng trong lĩnh vực thể thao như cầu thủ, huấn luyện viên, câu lạc bộ, các giải đấu … và bổ sung các quan hệ trọng điểm biểu diễn hoạt động trong thi đấu và chuyển nhượng bóng đá. Chúng là các thuộc tính liên kết các khái niệm trừu tượng trong ontology BKSport. Ví dụ, SportPerson, SportTeam, Defender, Forward, Goalkeeper, Midfielder, SportPerson move-to SportTeam, SportTeam sign-with SportTeam, SportTeam concern-with SportPerson, Coach buy Defender. Có thể nói ontology BKSport có ba tầng ngữ nghĩa: tầng trừu tượng và khái quát là từ vựng kế thừa từ PROTON, tầng trung gian tái sử dụng một phần từ ontology BBC Sport, và tầng chi tiết là các khái niệm và thuộc tính được thiết kế và bổ sung bởi tác giả.

Hình 2.10 mô tả một cách trực quan về một phần của ontology BKSport. Các hình ôvan biểu diễn các lớp, còn các cạnh có mũi tên biểu diễn cho một quan hệ giữa 2 lớp. Theo chiều mũi tên của một cạnh, lớp thứ nhất là "Domain" của quan hệ và lớp thứ hai là "Range" của quan hệ. Khi cạnh có mũi tên ở cả hai đầu, nghĩa là 2 lớp tương ứng với cạnh này vừa có thể đóng vai trò là "Domain", vừa có thể đóng vai trò là "Range". Các lớp như Team, Forward, Defender là các lớp chi tiết hơn cho lĩnh vực thể thao. Các lớp như Organization, SportEvent, Person là các lớp ở mức cao của PROTON.

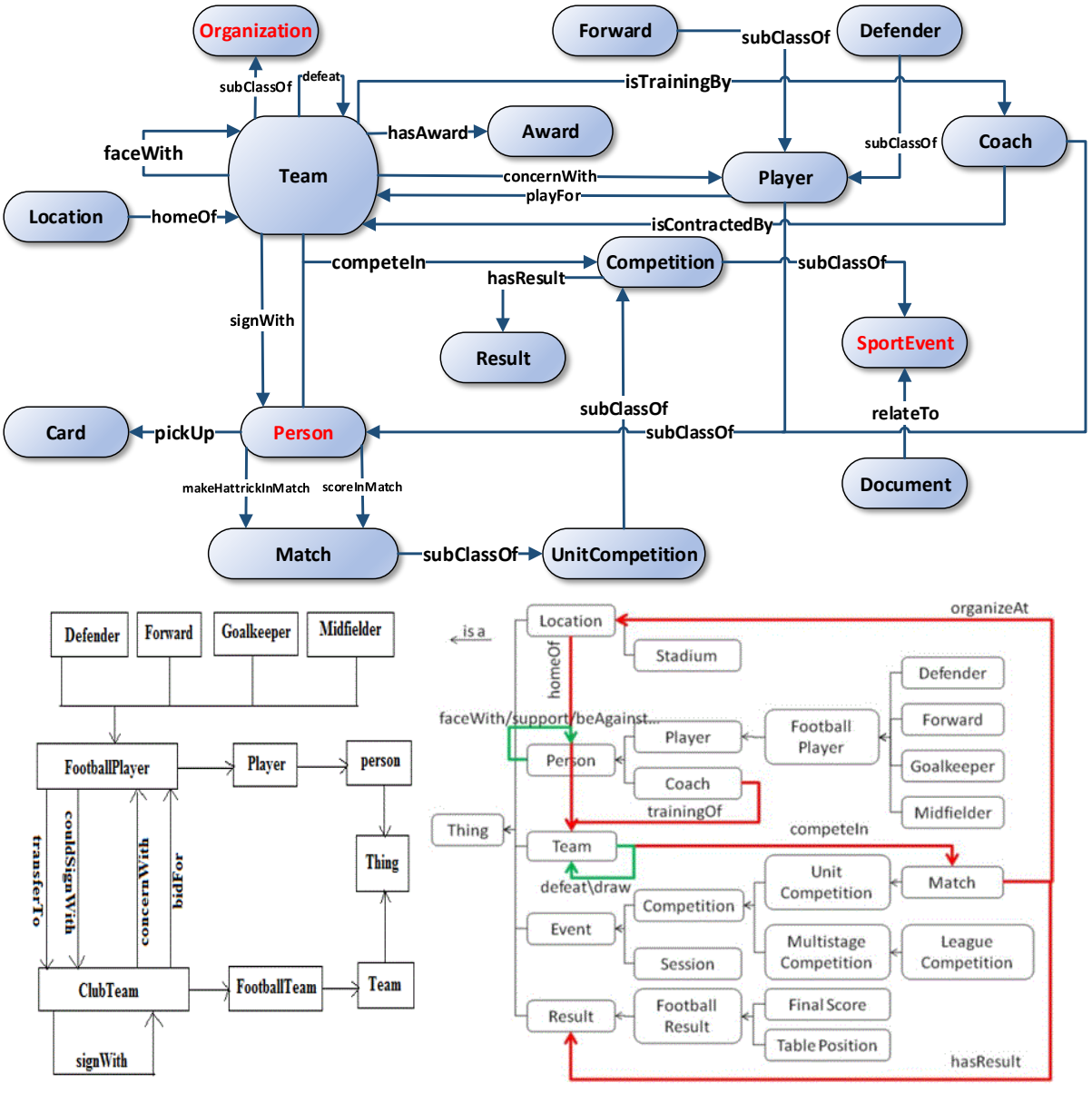
Dưới đây là mô tả ngắn gọn về một số lớp chính trong BKSport ontology:

- Team: bao gồm các thông tin về các đội bóng thi đấu ở các giải đấu trên thế giới. Các thông tin đó gồm tên đội bóng, quốc gia, số lượng cầu thủ, giải đấu …

- Player: bao gồm các thông tin mô tả về một cầu thủ như tên, tuổi, giới tính, đội bóng đang thi đấu … Nó là lớp cha của các lớp con chi tiết hơn như Forward, Defender. Các lớp con này chứa thêm các thông tin về vị trí thi đấu của cầu thủ trong đội bóng.
- Coach: bao gồm các thông tin mô tả về một huấn luyện viên như tên, tuổi, câu lạc bộ đang huấn luyện …
- Competition: bao gồm các thông tin mô tả về một trận thi đấu giữa hai đội bóng.
- Result: mô tả kết quả thi đấu giữa hai đội bóng.

Một số quan hệ chính giữa các lớp được liệt kê như sau:

- competeIn: Một đội bóng thi đấu trong một giải đấu.
- playFor: Thể hiện quan hệ rằng một cầu thủ đang chơi cho một câu lạc bộ.
- subClassOf: Một lớp có thể là lớp con của một lớp khác. Lớp con là lớp kế thừa các thông tin của lớp cha, đồng thời bổ sung các thông tin ở mức độ chi tiết hơn. Ví dụ, Forward là lớp con của (subClassOf) lớp Player. Ngoài các thông tin chung về một cầu thủ, lớp Forward còn cho biết rằng cầu thủ này chơi ở vị trí tiền đạo (chứ không phải ở một vị trí khác như hậu vệ hay thủ môn).
- hasResult: Một trận đấu có một kết quả thi đấu. Ví dụ: trận đấu giữa Chelsea và Liverpool (Competition) có kết quả (hasResult) là 1-1 (Result).



Hình 2.10 *Một phần của ontology BKSport*

2.3.3 Thu thập và tiền xử lý tin tức

Đầu tiên, dữ liệu từ các Website thể thao nổi tiếng như Sky Sports, ESPN sẽ được thu thập bởi thành phần Crawler. Sau đó, thành phần Preprocessor sẽ tiền xử lý dữ liệu thu được từ Crawler, loại bỏ các tin tức dư thừa (ví dụ, các nội dung quảng cáo) và giữ lại nội dung chính của tin tức. Các dữ liệu có ích như tiêu đề của tin tức, các liên kết có liên quan... cũng được giữ lại bởi vì những dữ liệu này có những ràng buộc với tin tức, vì thế việc phân tích chúng có thể giúp cho hệ thống nhận dạng và trích rút thông tin chính xác hơn.

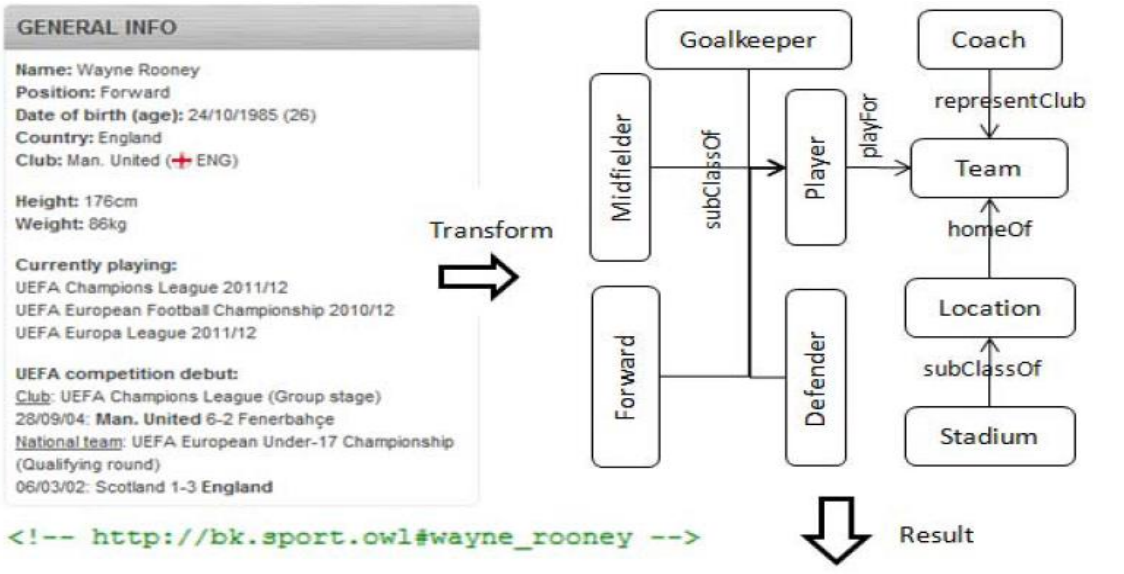
2.3.4 Xây dựng cơ sở tri thức thể thao

Để trích rút thông tin ngữ nghĩa, chúng ta cần phải có một cơ sở tri thức đủ lớn. Thành phần Web Scrapper thu thập cơ sở dữ liệu liên quan đến thể thao như cầu thủ (tên, tuổi, môn thể thao, …), câu lạc bộ (tên, sân nhà, ...), trận thi đấu, giải thưởng, sân vận động ... và gửi chúng đến thành phần làm giàu cơ sở tri thức (Knowledge Base Enrichment). Một mô đun con của nó – mô đun chuyển đổi dữ liệu tự động và sinh RDF – sẽ chuyển đổi tự động dữ liệu sang định dạng RDF.

Một mô đun khác của KBE chịu trách nhiệm nhập thủ công các bí danh của thực thể hoặc các từ đồng nghĩa mà không thể thu thập và chuyển đổi tự động. Dữ liệu RDF được chuyển đổi sau đó được nhập vào cơ sở tri thức thể thao của hệ thống.

Cho đến nay, luận án đã bổ sung cơ sở tri thức về các cầu thủ, các huấn luyện viên, các sân vận động bóng đá v.v. của giải đấu Premier League, La Liga, Champions League, các tay vợt tennis từ ATP rankings.

Hình 2.11 dưới đây mô tả một phần quá trình làm giàu cơ sở tri thức KIM với dữ liệu thể thao. Phần trên bên trái của hình 2.11 là một phần của trang Web HTML hiển thị dữ liệu về Wayne Rooney, phần trên bên phải của hình 2.11 là một phần của ontology thể thao BKSport đang biểu diễn lĩnh vực bóng đá và phần dưới hình 2.11 là cơ sở tri thức về Wayne Rooney sau khi được xử lý.




```
<!-- http://bk.sport.owl#wayne_rooney -->
```

Result

```
<owl:NamedIndividual rdf:about="http://bk.sport.owl#wayne_rooney">
    <rdfs:label xml:lang="en">Wayne Rooney</rdfs:label>
    <protons:mainLabel>Wayne Rooney</protons:mainLabel>
    <rdf:type rdf:resource="&bk;Forward"/>
    <protons:generatedBy rdf:resource="http://bk.sport.owl"/>
    <protons:hasMainAlias rdf:resource="&bk;rooney_alias"/>
    <protons:hasAlias rdf:resource="&bk;wayne_mark_rooney_alias"/>
    <playFor rdf:resource="&bk;manchester_united_fc"/>
</owl:NamedIndividual>
```

Hình 2.11 *Trích rút và xác định lớp ngữ nghĩa cho thực thể có tên*

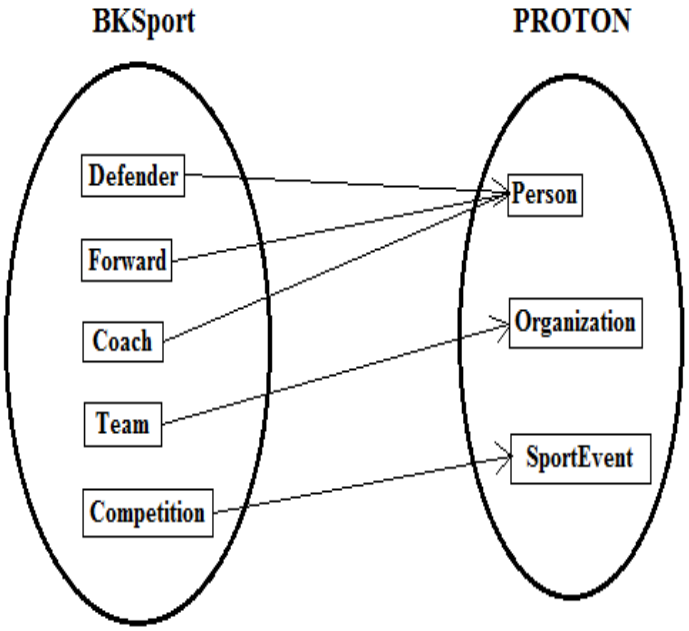
2.3.5 Nhận dạng, trích rút và xác định lớp ngữ nghĩa cho thực thể có tên

2.3.5.1 Nhận dạng thực thể có tên trong tin tức như là một thể hiện thuộc cơ sở tri thức

Để hiểu được ngữ nghĩa của văn bản, đầu tiên hệ thống cần hiểu được ngữ nghĩa của các thực thể có tên mà tên của chúng xuất hiện trong văn bản. Các thực thể có tên trong lĩnh vực thể thao bao gồm tên của các cầu thủ, các huấn luyện viên, các câu lạc bộ, các sân vận động, các sự kiện thể thao v.v. Ví dụ, đối với câu "Cordoba has completed the loan signing of Brazillian Winger Ryder Matos", hệ thống cần hiểu rằng Cordoba là tên của một câu lạc bộ bóng đá và Ryder Matos là tên của một Winger. Để làm điều này, phải có bước nhận dạng các thực thể có tên.

Sau khi đã được tiền xử lý, thông tin được chuyển đến thành phần nhận dạng thực thể có tên để phát hiện sự xuất hiện của cầu thủ, huấn luyện viên, câu lạc bộ, các tác nhân v.v trong các tin tức. Mô đun trích rút thực thể có tên lấy ra tất cả các thể hiện và các khái niệm của cơ sở tri thức mà xuất hiện trong các trang Web. Đóng vai trò này là tác vụ NER của hệ thống BKSport trong đó có tái sử dụng Ontology Proton của KIM.

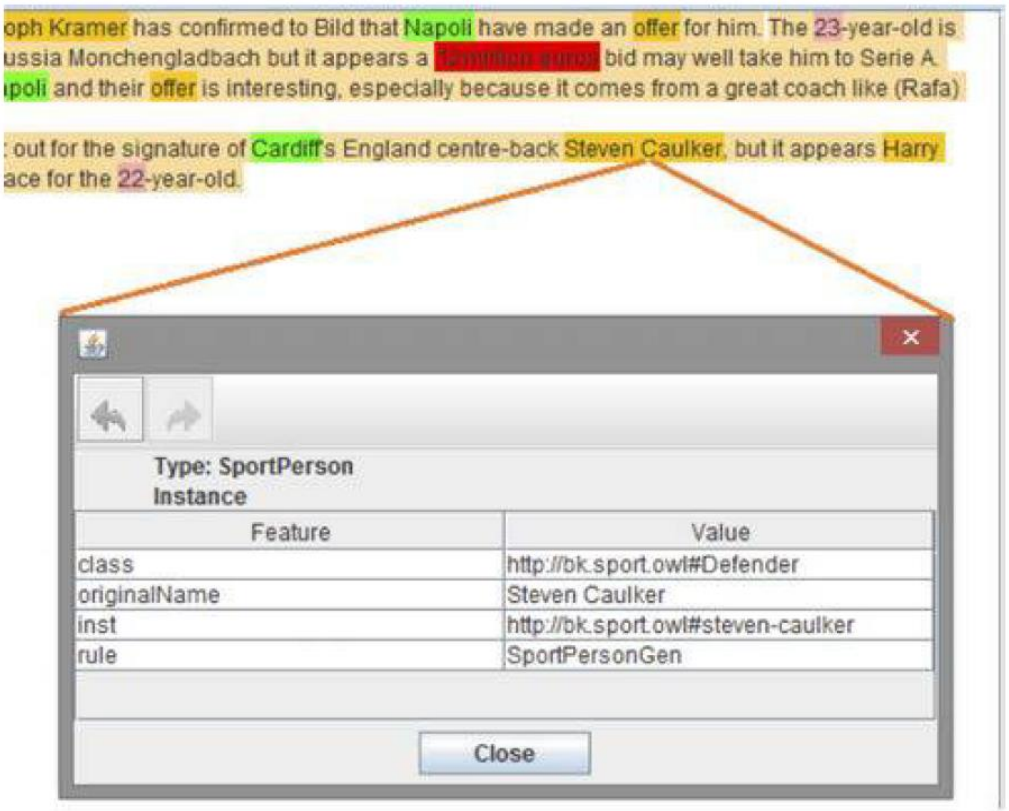
KIM [36] là một nền tảng mà luận án tái sử dụng để nhận dạng các thực thể có tên. KIM đã được xây dựng để nhận dạng các thực thể trong lĩnh vực tổng quát chung, nó không đặc thù cho một lĩnh vực cụ thể nào cả. Vì vậy để nhận dạng các thực thể ở mức sâu hơn và chi tiết hơn trong lĩnh vực thể thao, tác giả đã thêm một tập các khái niệm và các thuộc tính mới vào trong ontology của KIM, và bổ sung các thực thể mới vào cơ sở tri thức của KIM. Trong ontology mặc định của KIM (ontology PROTON), các thực thể có tên được biểu diễn ở mức khái quát (ví dụ, Person (người), Location (địa điểm)), không chi tiết (ví dụ, Winger, Forward). Do đó, tác giả đã tích hợp ontology BKSport với PROTON theo cách thức là các khái niệm cụ thể hơn của BKSport sẽ thay thế các khái niệm trừu tượng của PROTON trong quá trình nhận dạng. Nhờ tính mở của nền tảng KIM, việc tích hợp có thể được thực hiện bằng cách ánh xạ các khái niệm giữa chúng. Ví dụ, các lớp của ontology BKSport như Coach, Winger, Forward và Defender được hiểu như là các lớp con của lớp Person của PROTON. Hình 2.12 miêu tả một số lớp được ánh xạ từ ontology BKSport tới ontology PROTON.



Hình 2.12 *Một số ánh xạ từ BKSport đến PROTON*

Ánh xạ ontology không đảm bảo chắc chắn sự thành công của tác vụ NER khi không có sự bổ sung cơ sở tri thức về tin tức chuyển nhượng bóng đá. Để xây dựng cơ sở tri thức này, các cơ sở dữ liệu trên Web có chứa thông tin về các cầu thủ, các huấn luyện viên, các câu lạc bộ và

các tác nhân bóng đá trong các giải đấu bóng đá hàng đầu của châu Âu được thu thập và chuyển đổi thành chú thích ngữ nghĩa sử dụng ontology BKSport. Với việc mở rộng PROTON bằng ontology BKSport và sử dụng thư viện của nền tảng KIM, các thực thể có tên trong tin tức thể thao đã được nhận dạng đúng với lớp khái niệm định nghĩa trong ontology. Trong hình 2.13, Steven Caulker không chỉ được hiểu là Person mà còn được hiểu là một SportPerson, cụ thể hơn là một Defender.



Hình 2.13 *Nhận dạng thực thể có tên trong tin tức thể thao như một thể hiện của cơ sở tri thức*

2.3.5.2 Phát hiện bí danh của thực thể

Một đặc thù của lĩnh vực thể thao là các nhân vật, tổ chức nổi tiếng ngoài tên gọi chính thức thường hay có những biệt danh được biết đến rộng rãi bởi công chúng. Ví dụ huấn luyện viên Alex Ferguson có biệt hiệu Fergie, Lionel Messi gắn với "La Pulga", hay đội bóng đá FC Barcelona có biệt danh Barca hoặc Blaugrana. Do đó, việc phát hiện được các biệt danh này từ văn bản và ánh xạ chúng với các thực thể có tên đại diện chính thức tương ứng sẽ làm tăng hiệu quả của tác vụ nhận dạng thực thể có tên. Luận án thực hiện tác vụ này bằng cách tạo ra thông tin về các bí danh này khi xây dựng cơ sở tri thức thể thao một cách tự động (đã trình bày ở tiểu mục 2.3.4) sử dụng thuộc tính proton:hasAlias.

Khi các thông tin về các tên gọi khác của thực thể được bổ sung vào cơ sở tri thức BKSport dưới dạng bí danh (Alias) thì các thực thể này cũng được nhận dạng như thực thể chính.

2.3.5.3 Nhận dạng các thực thể ở mức khái niệm chi tiết

Mục tiêu của tác vụ này là phát hiện được các thực thể của các lớp chi tiết trong ontology BKSport như là "Defender" hay "Forward" thay vì các khái niệm ở mức cao như "Person" hay "Player". Qua phân tích nhận thấy, hầu hết các thực thể đều được biểu diễn dưới dạng "chức nghiệp" + "tên riêng" Ví dụ, "Striker Romelu Lukaku double leads Man United to easy win over CSKA Moscow". Ở đây công việc của nhân vật thể thao được nhận dạng là Striker. Các chức nghiệp thường chính là các nhãn của khái niệm, nên thuật toán sử dụng nhãn (label) của khái niệm làm mẫu (pattern) để xây dựng luật nhận dạng cho thực thể của từng khái niệm.

2.3.5.4 Cải tiến nhận dạng thực thể có tên ở dạng rút gọn

Trong các văn bản, thường sau khi sử dụng tên đầy đủ của thực thể, thực thể đó sẽ được nhắc lại với tên rút gọn để khiến bài viết trở nên ngắn gọn và dễ đọc (Ví dụ "Lionel Messi" được viết thành "Messi"). Bởi vậy, việc cải tiến để cung cấp khả năng nhận biết thực thể khi được biểu diễn với tên rút gọn rất quan trọng. Tên rút gọn thường sẽ là một phần của tên đầy đủ. Do đó, một thực thể khi được biểu diễn với tên rút gọn có thể được nhận biết khi nó đã được phát hiện với tên đầy đủ trước đó qua việc sử dụng phép toán so khớp một phần.

2.3.5.5 Nhận dạng thực thể cùng tên khác kiểu

Đây là trường hợp thường gặp trong chú thích văn bản khi thực thể có tên xuất hiện trong văn bản có thể thuộc về các kiểu khác nhau. Ví dụ, Santiago Bernabéu là tên một cầu thủ, nhưng cũng là tên một sân vận động. Ở trường hợp này, chúng ta sẽ tiến hành kiểm tra mẫu thực thể, tùy vào hậu tố theo sau để xác định kiểu của thực thể. Ở ví dụ trên nếu Santiago Bernabéu đi theo sau là khái niệm stadium thì thực thể bắt được sẽ được nhận dạng là sân vận động.

2.3.6 Trích rút "ngữ nghĩa" từ tin tức

Phát hiện và trích rút các ngữ nghĩa của thông tin là nội dung nghiên cứu quan trọng nhất để tạo ra chú thích ngữ nghĩa. Tác vụ này sử dụng kết quả từ giai đoạn nhận dạng thực thể có tên. Có nhiều khía cạnh ngữ nghĩa khác nhau được luận án quan tâm.

2.3.6.1 Các ngữ nghĩa bộ ba đơn giản

Trong tin tức thể thao, có một số ngữ nghĩa phổ biến ở dạng bộ ba <subject> <predicate> <object> diễn tả các sự kiện, hành động, kết quả … Ví dụ, tin tức thể thao có thể chứa "Barcelona won Arsenal", "Alex Ferguson defends Wayne Rooney", "Cristiano Ronaldo's transfer to Juventus" … Người dùng khi tìm đọc tin tức có thể muốn tìm kiếm các thông tin trên. Vì vậy, một trong những thuật toán đầu tiên được đề xuất là để phát hiện các ngữ nghĩa này.

Có ba mẫu trừu tượng chính mô tả ngữ nghĩa như sau:

a) <Person> <relation> <Person>. Ví dụ, <Marcus Rashford> <be against> <Jose Mourinho>.

b) <Organization> <relation> <Organization>. Ví dụ, <Manchester City> <defeat> <Arsenal> hoặc <Barcelona> <1:3> <Real Madrid>.

c) <Person> <relation> <Organization>. Ví dụ, <Romelu Lukaku> <transferTo> <Manchester United>.

Với mẫu đầu tiên, Person có thể là thực thể có tên như Marcus Rashford, Lionel Messi, hoặc khái niệm như Striker, Coach, hoặc đại từ như he, they. Quan hệ giữa các Person được nhận dạng bằng ontology, ví dụ <Person> <support> <Person>, <Person> <remind> <Person>. Một quan hệ có thể được mô tả bởi nhiều nhãn khác nhau tương ứng với các từ đồng nghĩa, ví dụ "surprise" và "stun" cùng mô tả quan hệ <surprise>.

Với mẫu thứ hai (<Organization> <relation> <Organization>), luận án tập trung vào kết quả của một trận đấu hoặc thông tin về một CLB đối đầu với một CLB khác.

Luận án sử dụng mẫu cuối cùng để trích rút thông tin về thái độ của cầu thủ/huấn luyện viên/trọng tài đối với một CLB/liên đoàn/giải đấu.

2.3.6.2 Ngữ nghĩa về thực thể quan trọng trong tin tức

Bên cạnh việc phát hiện các cặp bộ ba ngữ nghĩa đơn giản, luận án còn đề xuất thuật toán sinh chú thích ngữ nghĩa cho các thực thể có tên xuất hiện trong tin tức. Quan trọng hơn là những chú thích cho các thực thể liên quan đến những thông tin quan trọng trong tin tức. Nhiệm vụ này liên quan đến việc xác định các thực thể chính mà tin tức đề cập đến, bên cạnh việc tạo siêu dữ liệu cơ bản như các tiêu đề. Thuật toán định nghĩa một trọng số cho mỗi thể hiện để xác định xem nó có quan trọng trong mục tin hay không. Việc tính toán trọng số này dựa trên tần suất xuất hiện của mỗi thể hiện, vị trí xuất hiện của chúng trong văn bản, và mối quan hệ giữa

thể hiện với các khái niệm khác có trong ontology. Ngoài ra, khi áp dụng luật trích chọn, trọng số phụ thuộc giữa lớp của thể hiện cũng được so khớp với chính luật đó. Thuật toán trích rút các sự kiện đơn giản và các thực thể quan trọng trong tin tức được trình bày như sau:

Thuật toán 1: Sinh các chú thích ngữ nghĩa về thực thể quan trọng trong tin tức
Input: wc_c - weight of concept c for the news content
wt_c - weight of concept c for the news title
wd_c - distance weight of concept c with other concepts wr_c - weight of concept c with extraction rule r.
R - set of extraction rules, $W_{total} = 0$
Output: tập các bộ ba (triple) diễn đạt thông tin tin tức có tiêu đề là gì, liên quan đến các thực thể quan trọng nào.

Extract triple: <webpage.uri bk:hasTitle webpage.title>
for each named entity ***i*** recognized as instance of concept ***c***
m = number of occurences of i in title.
$W_{title-i} = m^* wt_c$
k = number of occurences of i in content.
$W_{content-i} = k^* (wc_c + wd_c)$, $W_{semantic-i} = 0$
foreach ***sen*** in {news sentences} do
foreach rule ***r*** in R do
compare ***r*** with *annotations* in ***sen***
if ***r*** matchs instance ***i***{
Extract triple corresponding ***r***
$W_{semantic-i} = W_{semantic-i} + wr_c$
endfor
endfor
$W_i = W_{title-i} + W_{content-i} + W_{semantic-i}$
$W_{total} = W_{total} + W_i$
endfor
meanW = W_{total} / number of entities
for each named entity ***i*** recognized in news
if W_i > meanW
Extract triple <webpage.uri bk:about element.uri.>
else Extract triple
<webpage.uri bk:contain element.uri.>
endfor

Ý tưởng chính của thuật toán đề xuất là: các luật được so khớp với nhãn chú thích được phát hiện bởi mô đun trích rút thực thể có tên (NEE) trong từng câu để tìm ra thông tin ngữ nghĩa. Vấn đề là trong số rất nhiều thực thể có tên được phát hiện, thực thể nào được quyết định là quan trọng với tin tức. Trọng số cuối cùng của một thể hiện tương ứng với một chú thích được đánh giá dựa trên số lần xuất hiện của nó trong tin tức và trọng số của luật so khớp. Mỗi thể hiện được lựa chọn là quan trọng đối với tin tức nếu trọng số của nó lớn hơn trọng số trung bình của tất cả các thể hiện.

Dữ liệu thông tin quan trọng của tin tức và tiêu đề chứa ý chính của tin tức. Trong rất nhiều trường hợp, ngữ nghĩa được trích rút từ tiêu đề là thông tin chính của tin tức. Do đó, luận án tập trung phân tích tiêu đề của tin tức. Mỗi thể hiện được nhận dạng trong tiêu đề có trọng số lớn hơn các thể hiện khác.

2.3.6.3 Chú thích ngữ nghĩa về tuyên bố gián tiếp

Bên cạnh các mối quan hệ thông thường, các tuyên bố gián tiếp cũng rất thường xuyên được đưa ra trong tin tức thể thao. Cũng tương tự như các quan hệ khác, quan hệ này cũng được nhận dạng dựa trên các mô hình được xây dựng từ tập từ khóa mô tả các quan hệ. Bảng 2.1 dưới đây mô tả các từ khóa và mô hình nhận dạng quan hệ này.

Bảng 2.1. Từ khóa cho các câu tuyên bố gián tiếp

Từ khóa	**Mô hình**
“say that”, “said that”, “announce”, “speech”	{SportPerson} [từ khóa] {Statement}
“statement”, “added”	{Statement}, {SportPerson} [từ khóa]

Từ các quan hệ được mô tả trong BKSport Ontology và các mô hình tương ứng với các quan hệ đó, tác giả đã sử dụng JAPE để xây dựng các luật nhận dạng quan hệ. Mỗi quan hệ sẽ có một luật tương ứng nhận dạng. Tuy nhiên, tất cả đều hoạt động theo một nguyên tắc chung: nếu một mô hình được tìm thấy, thì sẽ sinh quan hệ tương ứng.

Riêng đối với trường hợp nhận dạng câu tuyên bố gián tiếp, luận án đi sâu vào phân tích các mệnh đề gián tiếp theo sau "said that ", "announce". Việc nhận dạng và sinh các chú thích ngữ nghĩa trong trường hợp này được trình bày như sau:

Thuật toán 2: Sinh các chú thích ngữ nghĩa về tuyên bố gián tiếp
Input: P = {A “said that”/”announce B”};
//P là một mẫu tuyên bố gián tiếp (ví dụ, A “said that” B, A “announce” B…)
Output: Các bộ ba (triple) diễn đạt tuyên bố gián tiếp

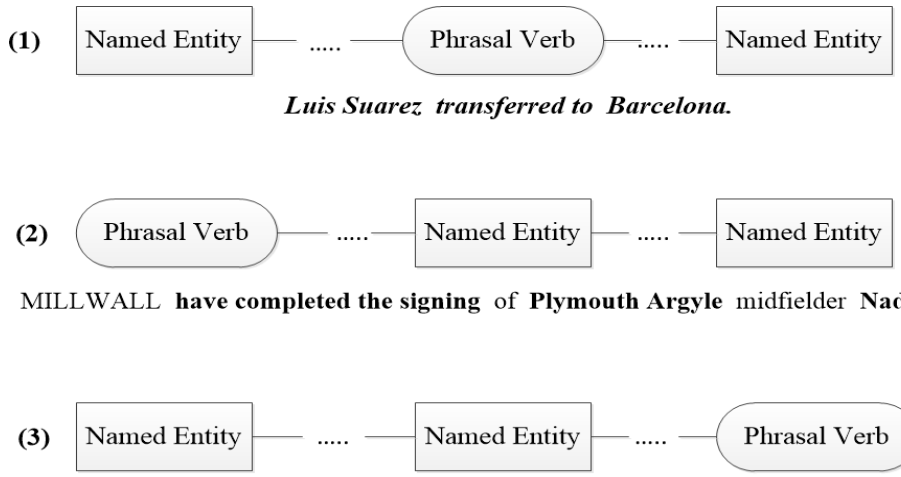
```
foreach (Chú_thích p trong P) do {
        statement = p.get(“B”);
        //chú thích các tuyên bố
        annotationSet = BKSport.annotate(statement);
        foreach (Annotation annotation in annotationSet){
        if (annotation.contains(“semantic”)) {
                //Tạo tuyên bố giống với chú thích
                subject=annotation.get(“subject”);
                predicate=annotation.get(“predicate”);
                object=annotation.get(“object”);
                //Sinh các bộ ba
                <A> <bksport:said that> <statement>;
                <statement> <rdf:subject>       subject;
                <statement> <rdf:predicate>     predicate;
                <statement> <rdf:object>        object;
                endif
        endfor
endfor
```

2.3.6.4 Chú thích ngữ nghĩa về tin tức chuyển nhượng

Trong thể thao, chuyển nhượng là một phân khúc tin tức hấp dẫn với các độc giả. Các tin tức về một cầu thủ chuyển từ câu lạc bộ này sang câu lạc bộ khác hoặc ký kết hợp đồng giữa hai câu lạc bộ đều được đăng tải trên nhiều nguồn tin tức khác nhau. Các chú thích ngữ nghĩa về tin tức trong chủ đề đặc thù này, nếu có thể được tạo ra sẽ làm phong phú thêm tập chú thích ngữ nghĩa của hệ thống BKSport và sẽ được khai thác bởi các chức năng của cổng tin tức ví dụ như giao diện tổng hợp tin tức chuyển nhượng, tìm kiếm ngữ nghĩa, liệt kê các tin tức liên quan. Tuy nhiên, chưa có nhiều nghiên cứu quan tâm đến vấn đề này. Không giống như thông tin về kết quả các trận đấu hoặc thông tin thể thao khác, thông tin chuyển nhượng bóng đá hàm chứa nhiều ngữ nghĩa đặc thù do đó việc trích rút chúng sử dụng mô hình bộ ba đơn giản khó đạt hiệu quả cao. Luận án đề xuất một phương pháp thích hợp để trích rút những ngữ nghĩa này, bổ sung vào kết quả chung của luận án về bài toán sinh chú thích ngữ nghĩa cho tin tức thể thao. Các kết quả nghiên cứu liên quan được tác giả trình bày trong bài báo "A novel approach for automatic extraction of semantic data about football transfer in sport news" tại tạp chí *International Journal of Pervasive Computing and Communications* (2015).

a) Một số mẫu nhận dạng quan hệ chuyển nhượng trong tin tức

Tin tức được diễn đạt bằng ngôn ngữ tự nhiên với các cấu trúc văn phạm và ngữ nghĩa đa dạng và phức tạp. Luận án không đặt mục tiêu tìm ra một tập các mô hình đại diện cho tất cả các ngữ nghĩa có thể về chuyển nhượng mà hướng tới việc xác định được các thành phần ngôn ngữ cấu thành nên những ngữ nghĩa quan trọng. Từ những khảo sát và nghiên cứu trên nhiều tin tức chuyển nhượng bóng đá để cố gắng tìm ra một số điểm chung về cấu trúc và các thành phần văn phạm của các ngữ nghĩa này, tác giả đi sâu phân tích mô hình bộ ba đơn giản để đề xuất ba mẫu nhận dạng ngữ nghĩa về chuyển nhượng như hình 2.14 sau:



Hình 2.14 *Các thành phần ngôn ngữ tự nhiên trong mẫu nhận dạng các quan hệ chuyển nhượng*

Các thành phần cơ bản cấu thành các mẫu nhận dạng trên bao gồm các thực thể có tên (named entity), cụm động từ (phrasal verb). Vì lĩnh vực đang được xem xét là lĩnh vực chuyển nhượng bóng đá, cho nên "thực thể có tên" thường chỉ là con người thể thao hoặc đội bóng. "Phrasal Verb" ở đây là cụm từ chứa "verb" + "adverb" hoặc "verb" + "preposition". Các động từ mô tả thuộc tính của các quan hệ chuyển nhượng, và "thời" của động từ sẽ xác định quan hệ thuộc về một trong ba trường hợp sau đây:

- Chuyển nhượng đã xảy ra
- Chuyển nhượng có thể xảy ra trong tương lai gần, và
- Chuyển nhượng đã không thành công

"Thời" của động từ phụ thuộc vào dạng của động từ hoặc phụ thuộc vào những từ mang ý nghĩa và đứng trước động từ. Trong ví dụ: "Former Rangers goalkeeper Scott Gallacher has

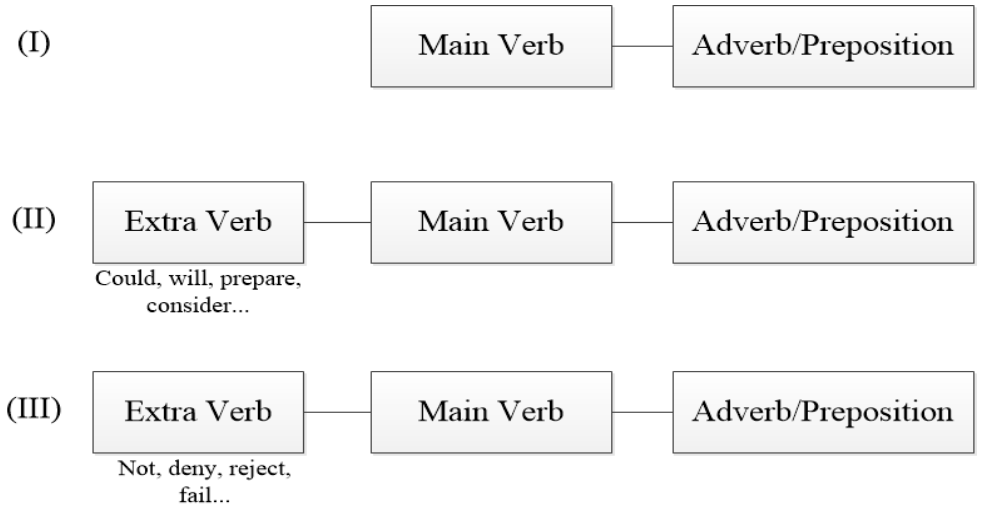
signed a two-year deal at Hearts", động từ "signed" cho thấy rằng chuyển nhượng đã xảy ra. Một ví dụ khác: "Barcelona forward Messi will make a new contract". Từ "will" đứng trước động từ "make" cho thấy rằng trường hợp này chuyển nhượng chưa xảy ra nhưng có thể xảy ra trong tương lai gần.

Cụ thể hơn, trong hình 2.15 bên dưới, luận án trình bày mô hình nhận dạng cụm động từ như sau:

<Extra Verb> <Main Verb> <Adverb/Preposition>

Trong đó:

- "Extra Verb" (trợ động từ) là các từ đứng ngay trước động từ chính, cho biết hành động hoặc sự kiện chuyển nhượng rơi vào một trong ba trường hợp sau: (1) sự kiện chưa xảy ra nhưng có thể xảy ra trong tương lai gần, (2) sự kiện đã xảy ra, và (3) sự kiện không xảy ra.
- "Main Verb" là động từ chính của "phrasal verb".
- "Adverb/Preposition" là trạng từ và giới từ kế tiếp làm thay đổi động từ.



Hình 2.15 *Các mẫu biểu diễn cụm động từ*

Nếu không có "extra verb" trước "main verb", tác giả cho rằng sự kiện đã xảy ra (ngữ nghĩa tích cực).

Nếu có "extra verb" trước "main verb", có hai trường hợp xảy ra: ngữ nghĩa phủ định và ngữ nghĩa đại diện cho các khả năng:

- "extra verb" mang ý nghĩa tương lai gần, cho biết hành động hoặc sự kiện có thể xảy ra trong tương lai, ví dụ "could", "prepare", "will", "consider".
- "extra verb" mang ý nghĩa phủ định, cho biết hành động hoặc sự kiện đã không và sẽ không xảy ra, ví dụ "not", "no", "don't", "fail", "reject".

b) Quy trình nhận dạng ngữ nghĩa quan hệ chuyển nhượng

Để thực hiện nhận dạng ngữ nghĩa quan hệ chuyển nhượng, các luật nhận dạng để trích rút và nhận biết các quan hệ ngữ nghĩa chuyển nhượng đã được thiết kế. JAPE [116] được lựa chọn là ngôn ngữ biểu diễn luật bởi vì nó có nhiều ưu điểm. JAPE là một thành phần của GATE, dùng để nhận dạng các thực thể được xác định bằng các luật, nó là ngôn ngữ được sử dụng để viết nên các biểu thức chính quy thông qua các chú thích.

Đầu tiên, văn bản được chia thành các câu, mỗi câu mang một nội dung nào đó. Các câu thường bắt đầu và kết thúc bởi dấu câu như dấu chấm ".", dấu chấm phảy ";" hoặc từ cho biết sự bắt đầu của nội dung mới như "while", "however", "but", vì thế ta có thể dùng các luật để dễ dàng thực hiện điều này. Sau đó, mỗi câu sẽ được đem so khớp với một danh sách các luật.

Luận án chỉ xem xét các thực thể có tên và các cụm động từ, vì vậy các từ không liên quan sẽ bị bỏ qua. Đối với các cụm động từ, bởi vì một động từ có thể ở trong nhiều dạng khác nhau

và nhiều từ khác nhau có thể diễn tả cùng một loại quan hệ ngữ nghĩa (ví dụ, “move to”, “big moves”, “transferred to” tất cả đều diễn tả quan hệ “bksport:transferTo”), cho nên trong khi định nghĩa các luật, các động từ có liên quan cũng được tập hợp vào trong các tập từ vựng. Ví dụ, một tập từ vựng biểu diễn việc ký kết được định nghĩa như sau:

```
Macro: SIGN
(
{Token.string=="sign"}|{Token.string=="signs"}|{Token.string=="signed"}|{Token.string=="signing"}|{
Token.string=="signature"}
)
```

Dưới đây là 2 phần của hai luật nhận dạng, Sign01 và Transfer01:

```
Rule:Sign01
Priority:80
(
        ({SportPerson}):p1
        ({Token.string!=".". Token.string!=";", !SportPerson})*
        (SIGN)
        ({Token.string!=".", Token.string!=";", Token.string!=","})*
):sign
Rule: Transfer01
Priority: 70
(
        ({SportPerson})p
        ({Token.string!=".", !SportPerson})*
        (TRANSFER)
        ({Token.string!=".", !SportPerson})*
        ({SportTeam}:t
):transfer
```

Để nhận dạng trong hai trường hợp này: sự kiện chuyển nhượng sẽ xảy ra trong tương lai gần và sự kiện chuyển nhượng không xảy ra, luận án dựa vào những mô hình đã được xây dựng. Theo đó, ngoài nhận dạng động từ chính trong tin tức như đã trình bày, luận án cũng phải nhận dạng “extra verb”. Tương ứng với hai trường hợp trên, luận án tạo ra hai tập từ vựng “extra verb”. Tập thứ nhất chứa các từ/cụm từ biểu diễn sự kiện sẽ xảy ra trong tương lai gần:

```
Macro: COULD
({Token.string=="could"} | {Token.string=="will"} | {Token.string=="prepare"} |
{Token.string=="consider"} | […])
```

Tập thứ hai chứa từ/cụm từ mà biểu diễn sự kiện không xảy ra:

```
Macro: NOT
({Token.string=="not"} | {Token.string=="deny"} | {Token.string=="reject"} |
{Token.string=="fail"} | […])
```

Sau đây là hai luật đơn giản Sign02 và Sign03 để nhận dạng các ngữ nghĩa mà thuộc về hai trường hợp nêu trên:

```
Rule: CouldSign01
Priority: 90
(
        ({SportPerson}):p1
        ({Token.string!=".", Token.string!=";", !SportPerson})*
        (COULD)
        ({Token.string!=".", Token.string!=";", !SportPerson})*
        (SIGN)
        ({Token.string!=".", Token.string!=";", Token.string!=","})*
        ({SportPerson}):p2
):couldsign
```

```
Rule: NotSign01
Priority: 100
(
        ({SportPerson}):p1
        ({Token.string!=".", Token.string!=";", !SportPerson})*
        (NOT)
        ({Token.string!=".", Token.string!=";", !SportPerson})*
        (SIGN)
        ({Token.string!=".", Token.string!=";", Token.string!=","})*
        ({SportPerson}):p2
):notsign
```

Gặp trường hợp một đoạn văn bản khớp với nhiều luật, luận án sẽ xử lý để chọn ra một luật phù hợp nhất theo các nguyên tắc sau:

- Nếu nhiều luật đều khớp với một vùng tài liệu bắt đầu tại một điểm X, luật nào khớp với vùng dài nhất sẽ được chọn. Ví dụ, với hai luật nêu ở trên (Sign01 và Transfer01), giả sử chúng ta có văn bản "Alexis Sanchez signed a contract with Pep Guardiola to move to Manchester City in the next season", thì luật Transfer01 sẽ được áp dụng bởi vì nó khớp với một vùng văn bản dài hơn bắt đầu từ cùng một điểm "Alexis Sanchez signed a contract with Pep Guardiola to move to Manchester City in the next season". Trong khi đó, luật Sign01 chỉ khớp với "Alexis Sanchez signed a contract with Pep Guardiola".
- Nếu nhiều luật đều khớp với một vùng tài liệu và có cùng độ dài, thì luật có độ ưu tiên cao hơn sẽ được chọn (luận án gán cho mỗi luật một giá trị độ ưu tiên nhất định, ví dụ, với hai luật trên, độ ưu tiên của luật Sign01 là 80 và độ ưu tiên của luật Transfer01 là 70).
- Nếu nhiều luật có cùng độ ưu tiên, luật nào được định nghĩa trước nhất sẽ được chọn.
- Nếu tất cả các điều kiện nêu trên đều như nhau, thì một luật nào đó sẽ được chọn ngẫu nhiên.

Cuối cùng, các luật sẽ ánh xạ những quan hệ được nhận dạng vào quan hệ tương ứng trong ontology để sinh biểu diễn RDF.

c) Chú thích các đại từ và cụm bí danh đặc biệt

Trong các văn bản dài, để tránh việc phải nhắc lại tên của các thực thể nhiều lần, người ta thường dùng các đại từ để thay thế. Điều này trực tiếp gây khó khăn đến việc nhận dạng các quan hệ ngữ nghĩa, vì để nhận dạng được các quan hệ, các thực thể có tên phải được nhận dạng.

Có một số nghiên cứu xoay quanh vấn đề nhận dạng đại từ. [117] đã thực hiện một nghiên cứu mô tả một thực hiện độc lập mà đã được công bố rộng rãi bởi Resolution of Anaphorea Procedure (RAP) do Lappin và Leass xây dựng. Nó xử lý các đại từ chỉ người ngôi thứ ba, các trùng lặp từ vựng, và nhận dạng các đại từ dư thừa (pleonastic pronouns) trong ngôn ngữ tiếng Anh. Nó đạt được độ chính xác là 57,9% với khuôn dạng đầu vào MUC-6.

Trong [118], các tác giả đã đề xuất một hệ thống giải quyết dư thừa có tên "Automatic Pronominal Anaphora Resolution in English Texts" dựa trên WordNet ontology và các luật Heuristic. Hệ thống được đề xuất này có khả năng giải quyết hiện tượng trùng lặp trong liên câu và trong nội dung câu trong văn bản tiếng Anh bằng một xử lý thích hợp với các đại từ dư thừa. Hệ thống đạt được tỉ lệ thành công tổng thể là 77%.

Nhằm nâng cao hiệu quả sinh chú thích ngữ nghĩa về hoạt động chuyển nhượng, luận án đề xuất một phương pháp để trích rút các đại từ và các cụm bí danh đặc biệt, dựa vào các luật trích rút thông tin. Chúng rất thích hợp khi áp dụng vào lĩnh vực thể thao. Tập các luật của luận án được xây dựng để biểu thị các đại từ phải tuân thủ những nguyên tắc sau đây:

- Các đại từ như 'he', 'him', 'I', 'me' đại diện cho SportPerson. Những đại từ như 'they', 'them', 'we', 'us' đại diện cho SportTeam.
- Các đại từ 'I', 'me', 'we', 'us' xuất hiện trong câu tuyên bố gián tiếp, đại diện cho tác nhân SportPerson hay SportTeam mà tuyên bố câu đó. Có hai mẫu câu nói gián tiếp như sau:

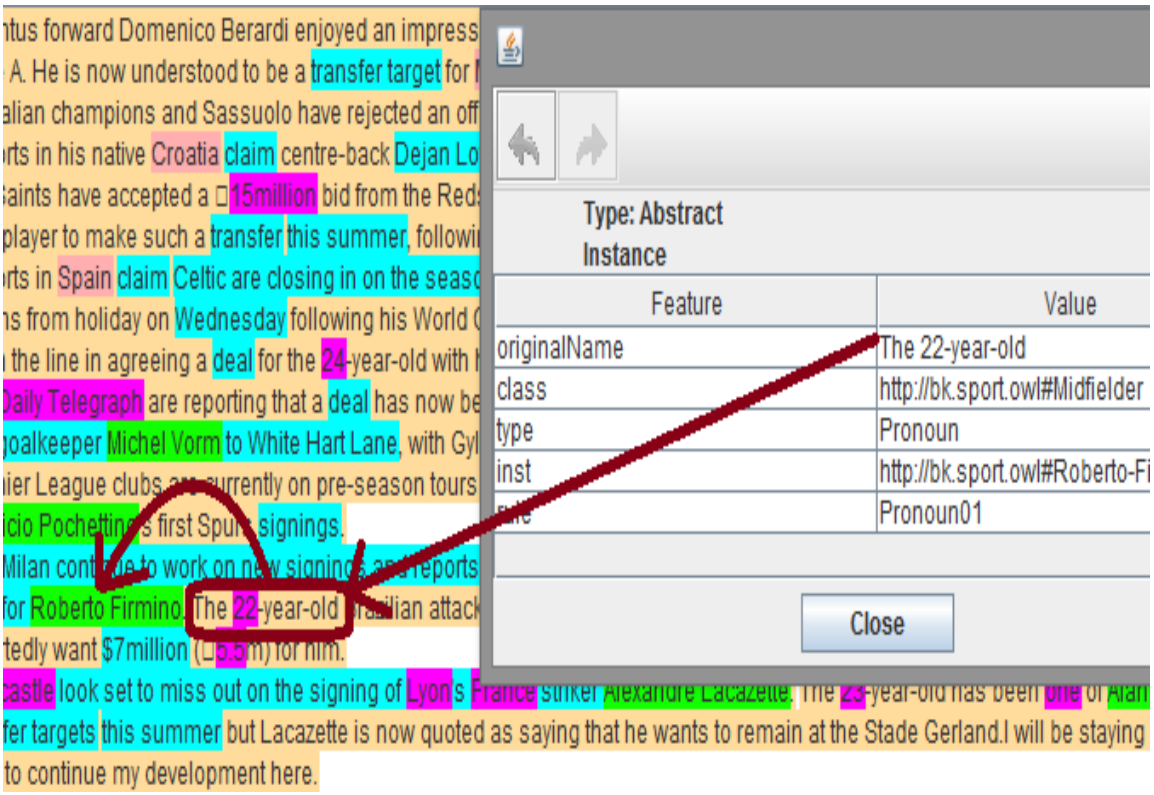
- tác nhân đứng trước câu nói gián tiếp.
- tác nhân đứng sau câu nói gián tiếp.

- Các đại từ đại diện cho những thực thể có tên SportPerson hoặc SportTeam mà xuất hiện trước và gần với đại từ đó. Trong trường hợp câu tuyên bố gián tiếp thì đại từ có thể đại diện cho các thực thể phía sau nó.
- Sau khi nhận dạng được các đại từ, luật này sẽ đặt lại trường class của các đại từ vào trong trường class của thực thể mà nó đại diện, để hỗ trợ cho việc nhận dạng các quan hệ chuyển nhượng.

Bên cạnh đó, các tin tức chuyển nhượng cũng thường xuyên sử dụng các cụm từ đặc biệt khác để biểu diễn các thực thể có tên. Ví dụ như dùng <'the' + number-year-old> để biểu diễn các cầu thủ được nhắc tới trước đó. Xem xét tin tức sau:

"Inter Milan continue to work on new signings and reports in Italy claim there has been contact with Bundesliga side Hoffenheim regarding a deal for **Roberto Firmino**. *The 22-year-old* Brazilian attacking midfielder has previously been linked with the likes of Liverpool, and Hoffenheim reportedly want $7million (£5.5m) for him".

Trong tin này, cụm từ "The-22-year-old" được dùng để thay thế cho cầu thủ Roberto Firmino, được thể hiện trong hình 2.16.



Hình 2.16 *Ví dụ về kết quả nhận dạng đại từ*

2.4 Thực nghiệm

Để đánh giá phương pháp đề xuất, luận án đã tiến hành thực nghiệm các thuật toán sinh chú thích ngữ nghĩa trên một tập tin tức thể thao được thu thập từ nhiều nguồn. Do phương pháp tổng thể là kết quả của nhiều nghiên cứu cho những bài toán con như nhận dạng và gán lớp cho các thực thể có tên, trích xuất chú thích ngữ nghĩa khác nhau ở dạng bộ ba, các kết quả thực nghiệm được trình bày theo thứ tự của các nghiên cứu này.

Tất cả các thực nghiệm được thực hiện trên máy tính Intel Core i7, CPU 2.30 GHz với RAM 8GB, hệ điều hành Microsoft Windows Server 2008. Các thuật toán được cài đặt bằng ngôn ngữ lập trình Java, sử dụng nền tảng KIM phiên bản 3.0.4.

Tập dữ liệu thực nghiệm

Trong giai đoạn đầu nghiên cứu chủ đề này, luận án đã thực hiện thực nghiệm trên một tập tin tức bóng đá của Giải bóng đá Ngoại hạng Anh (Premier League) và Giải bóng đá vô địch các câu lạc bộ châu Âu (Champions League). Hệ thống thu thập các tin tức từ nhiều nguồn nổi tiếng như skysports.com, premierleague.com với số lượng 150 tin tức (75 tin tức về Giải bóng đá Ngoại hạng Anh và 75 tin tức về Giải bóng đá vô địch các câu lạc bộ châu Âu).

Kịch bản thực nghiệm

Các thuật toán sẽ được thực thi trên mỗi tin tức trong tập dữ liệu thực nghiệm. Kết quả thu được sẽ được so sánh với kết quả của việc thực hiện các tác vụ tương ứng một cách thủ công bằng con người để xác định kết quả của thuật toán là chính xác hay không chính xác.

Để đánh giá hiệu quả của thuật toán nhận dạng thực thể có tên và thuật toán sinh chú thích, luận án sử dụng hai tham số tiêu chuẩn: độ chính xác (precision) và độ bao phủ (recall). Độ bao phủ (R) được xác định là tỉ lệ của kết quả chính xác thu được bởi thuật toán (RR) trên tổng số các kết quả chính xác cần được xác định (TRE). Độ chính xác (P) được xác định là tỉ lệ của kết quả chính xác thu được bởi thuật toán (RR) trên tổng số các kết quả nhận dạng mà thuật toán đưa ra (TR).

$$Precision(P) = \frac{Relevant\ recognized\ instances\ (triples)\ (RR)}{Total\ regconized\ instances\ (TR)} \quad (2.1)$$

$$Recall\ (R) = \frac{Relevant\ recognized\ instances\ (RR)}{Total\ relevant\ instances\ (TRE)} \quad (2.2)$$

Kết quả có thể được tính là các thực thể được nhận dạng, hoặc các bộ ba ngữ nghĩa được trích rút, tùy theo nội dung cần đánh giá. Thực nghiệm đánh giá phương pháp đề xuất trên hai tác vụ là:

- Phát hiện thực thể có tên trong tin tức thể thao.
- Phát hiện và trích rút ngữ nghĩa trong tin tức thể thao.

2.4.1 Nhận dạng thực thể có tên trong tin tức

Thực nghiệm này có mục đích đánh giá khả năng phát hiện thực thể có tên trong tin tức và gán chúng với các lớp trong ontology thể thao của phương pháp đề xuất trong luận án. Thuật toán đầu tiên luận án đề xuất có khả năng mở rộng KIM với cơ sở tri thức thể thao và nhận dạng được bí danh. Kết quả thực nghiệm được trình bày trong bảng 2.2.

Bảng 2.2. *Độ chính xác (P) và độ bao phủ (R) của quá trình trích rút từ 150 tin tức thể thao*

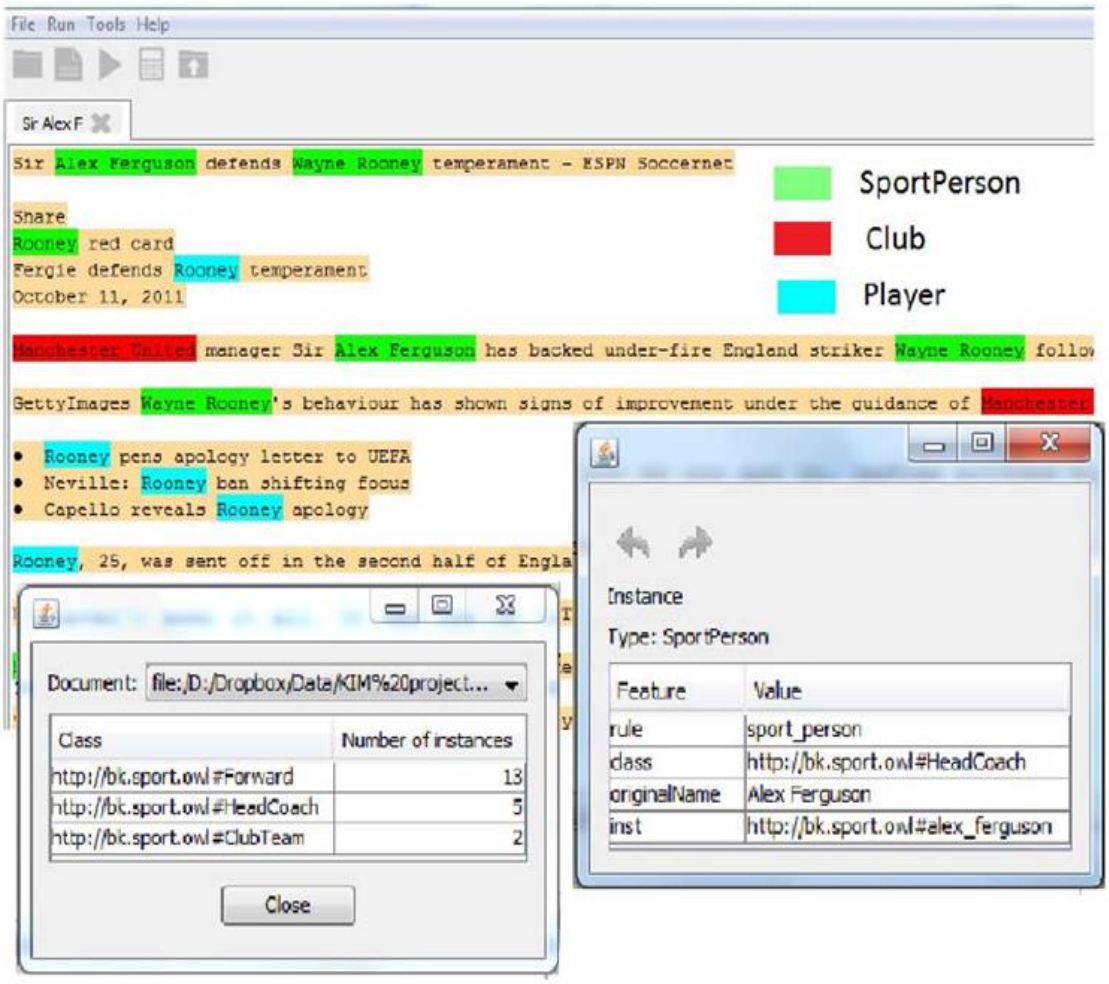
	TR	**RR**	**TRE**	**P%**	**R%**
Premier League	2018	1960	2674	97.1	73.3
Champion League	1240	1175	2590	94.7	45.3

Những số liệu được cung cấp ở bảng 2.2 cho thấy thuật toán đạt độ chính xác cao khi phát hiện tên của cầu thủ và câu lạc bộ nhờ có thành phần cơ sở tri thức thể thao được chuyển đổi tự động. Hình 2.18 cho thấy Wayne Rooney được trích rút như là một thể hiện của lớp Forward của Ontology thể thao thay vì lớp Person của PROTON như KIM đã làm. Nó chính xác và cụ thể hơn một thể hiện của khái niệm Person như hệ thống KIM đã đưa ra cho thực thể có tên Wayne Rooney.

Phương pháp cũng có khả năng nhận dạng nickname, bí danh của một số thực thể có tên. Ví dụ, Fergie được định nghĩa là nickname của Sir Alex Fersuson, trong khi Swans được nhận

dạng là bí danh của CLB bóng đá Swansea như được mô tả trong hình 2.18 và hình 2.20. Tuy nhiên, độ bao phủ vẫn chưa cao như mong đợi, với nguyên nhân được xác định là:

- Thiếu từ đồng nghĩa, các bí danh trong cơ sở tri thức. Những từ vựng này được nhập vào hệ thống theo cách thủ công và đòi hỏi nhiều thời gian để hoàn thành.
- Tập các luật trích rút chưa đầy đủ.



Hình 2.17 *Giao diện phần mềm sinh chú thích ngữ nghĩa*

Để tiến hành thực nghiệm và phát triển thành phần phần mềm sinh chú thích ngữ nghĩa cho hệ thống BKSport, luận án đã cài đặt các thuật toán trên trong một phần mềm chạy dưới dạng ứng dụng cũng như dưới dạng dịch vụ phần mềm. Phần mềm với giao diện đồ họa cho phép xem kết quả của thuật toán trên từng tin tức, phục vụ thực nghiệm trong khi dịch vụ chú thích ngữ nghĩa được triển khai kết nối với các thành phần khác trong hệ thống tổng hợp tin tức.

Hình 2.17 thể hiện giao diện đồ họa người dùng phần mềm phát triển bằng ngôn ngữ Java, minh chứng khả năng nhận dạng thực thể có tên như là thể hiện của các khái niệm (lớp) của ontology thể thao BKSport.

Hình 2.18 và hình 2.20 cho thấy các thể hiện được phát hiện trong hai tin tức thể thao "Fergie defends Rooney Temperament" và "Adebayor double help Spurs beat Swans" dựa trên nền tảng KIM và dựa trên phương pháp luận án đề xuất. Hình 2.19 và 2.21 minh họa các chú thích ngữ nghĩa tương ứng ở định dạng RDF được sinh ra bởi các thuật toán phát hiện bộ ba ngữ nghĩa đơn giản và xác định chủ đề của tin tức.

Hình 2.18 *Các thể hiện được nhận dạng bởi KIM và phương pháp đề xuất*

```
<owl:NamedIndividual rdf:resource="http://bk.sport.owl#news_0000D">
    <rdf:type rdf:resource="http://bk.sport.owl#News"/>
    <bk:hasUrl>Sir%20Alex%20Ferguson%20defends%20Wayne%20Rooney%20temperament%20
    -%20ESPN%20Soccernet.htm</bk:hasUrl>
    <bk:hasTitle xml:lang="en"> Fergies defends WayneRooney temperament</bk:hasTitle>
    <bk:about rdf:resource="http://bk.sport.owl#wayne_rooney"/>
    <bk:about rdf:resource="http://bk.sport.owl#alex_ferguson"/>
    <bk:contain rdf:resource="http://bk.sport.owl#manchester_united_fc"/>
    <bk:hasRelation rdf:resource="http://bk.sport.owl#defend_00001"/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="http://bk.sport.owl#defend_00001">
    <rdf:type rdf:resource="http://bk.sport.owl#Defend"/>
    <bk:hasAgent rdf:resource="http://bk.sport.owl#alex_ferguson"/>
    <bk:hasObject rdf:resource="http://bk.sport.owl#wayne_rooney"/>
</owl:NamedIndividual>
```

Hình 2.19 *Chú thích ngữ nghĩa được sinh ra với tin tức ở hình 2.18*

Hình 2.20 *Các thể hiện được nhận dạng bởi KIM và phương pháp đề xuất*

```
<owl:NamedIndividual rdf:resource="http://bk.sport.owl#news_0000E">
  <rdf:type rdf:resource="http://bk.sport.owl#News"/>
  <bk:hasUrl> adebayor-double-helps-spurs-beat-swans.html</bk:hasUrl>
  <bk:hasTitle xml:lang="en">Adebayor double helps Spurs beat Swans </bk:hasTitle>
  <bk:about rdf:resource="http://bk.sport.owl#tottenham_hotspur_fc"/>
  <bk:about rdf:resource="http://bk.sport.owl#rafael_van_der_vaart"/>
  <bk:about rdf:resource="http://bk.sport.owl#emmanuel_adebayor"/>
  <bk:contain rdf:resource="http://bk.sport.owl#premier_league"/>
  <bk:hasRelation rdf:resource="http://bk.sport.owl#defeat_00004"/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="http://bk.sport.owl#defeat_00004">
  <rdf:type rdf:resource="http://bk.sport.owl#Defeat"/>
  <bk:hasAgent rdf:resource="http://bk.sport.owl#tottenham_hotspur_fc"/>
  <bk:hasObject rdf:resource="http://bk.sport.owl#swansea_city_fc"/>
  <bk:hasScore xml:datatype="http://www.w3.org/2001/XMLSchema#string">
  3-1</bk:hasScore>
</owl:NamedIndividual>
```

Hình 2.21 *Chú thích ngữ nghĩa được sinh ra với tin tức ở hình 2.20*

2.4.2 Trích rút ngữ nghĩa từ tin tức thể thao

Thực nghiệm 1: Trong lần thực nghiệm đầu tiên, luận án đánh giá hiệu quả của thuật toán sinh chú thích ngữ nghĩa cho tin tức thể thao ở phiên bản thứ nhất. Tại thời điểm này thuật toán cho phép phát hiện ngữ nghĩa ở dạng bộ ba đơn giản và các ngữ nghĩa về chủ đề mà tin tức liên quan tới. Các kết quả được trình bày trong bảng 2.3, trong đó TG, CT và ET tương ứng là số bộ ba được sinh ra bởi thuật toán, số bộ ba sinh ra bởi thuật toán được xác định là đúng và số bộ ba được tạo ra bởi con người.

Bảng 2.3. *Kết quả trích rút thông tin ngữ nghĩa của thực nghiệm 1*

	TG	CT	ET	P%	R%
Premier League	213	198	313	92.9	68.0
Champion League	177	150	252	84.7	59.5

Thực nghiệm 2: Thực nghiệm này được tiến hành tại thời điểm luận án đã đề xuất thêm phương pháp sinh ra các chú thích về tuyên bố gián tiếp, cũng như tiến hành một số cải tiến nhằm nâng cao hiệu quả của tác vụ nhận dạng thực thể có tên bao gồm: nhận dạng các thực thể ở mức khái niệm chi tiết, nhận dạng thực thể có tên ở dạng rút gọn, nhận dạng thực thể cùng tên khác kiểu.

Hình 2.22 dưới đây minh họa kết quả chú thích ngữ nghĩa về các tuyên bố gián tiếp được sinh ra bởi thuật toán ở lần cải tiến này.

Hình 2.22 *Chú thích ngữ nghĩa về tuyên bố gián tiếp được trích rút*

Bảng 2.4 trình bày hiệu quả của thuật toán trong nhận dạng thực thể có tên, cũng như tạo ra các chú thích ngữ nghĩa bao gồm chú thích bộ ba đơn giản, chú thích về chủ đề tin tức, chú thích về tuyên bố gián tiếp. Có thể thấy rằng, với những cải tiến đã thực hiện, giá trị độ chính xác (P) và độ bao phủ (R) thu được cao hơn so với thực nghiệm đầu tiên, cả trong phát hiện thực thể có tên và sinh chú thích ngữ nghĩa. Ngoài ra, những bộ ba phức tạp như tuyên bố gián tiếp bây giờ đã được nhận dạng và sinh chú thích. Đây là kết quả của việc áp dụng những mô hình và luật do tác giả đề xuất. Tuy nhiên, độ bao phủ (R) vẫn cần được cải thiện bởi vì khối

lượng dữ liệu trong cơ sở tri thức của BKSport vẫn chưa đủ lớn và phong phú. Bên cạnh đó, số lượng các mẫu luật trích chọn để nhận ra các quan hệ là chưa đủ để bao phủ mọi trường hợp.

Bảng 2.4. *Thống kê nhận dạng thực thể có tên và bộ ba của thực nghiệm 2*

	TR	**RR**	**TRE**	**P%**	**R%**
Named Entities Recognition	2699	2692	4415	99,74	60,97
Triples Extraction	1002	890	1663	88,82	53,52

Thực nghiệm 3: Đánh giá hiệu quả của thuật toán sinh chú thích cho tin tức chuyển nhượng.

Đối với bài toán sinh chú thích ngữ nghĩa tin tức chuyển nhượng, tập dữ liệu thực nghiệm được mở rộng lên 237 tin tức chuyển nhượng được lấy từ nguồn Sky Sports. Thông qua tác vụ chú thích ngữ nghĩa thủ công trên tập dữ liệu này, 264 bộ ba ngữ nghĩa liên quan đến chuyển nhượng bóng đá đã được xây dựng. Luận án tiến hành thử nghiệm phương pháp trong hai kịch bản:

- Không sử dụng các luật nhận dạng đại từ.
- Sử dụng các luật nhận dạng đại từ.

Bảng 2.5 trình bày những kết quả thực nghiệm thu được từ lần thực hiện phiên bản đầu tiên của thuật toán. Số liệu cho thấy, việc sử dụng các luật nhận dạng đại từ giúp nâng cao hiệu quả của phương pháp.

Bảng 2.5. *Kết quả bước đầu của thực nghiệm nhận dạng quan hệ ngữ nghĩa*

	TRE	**TR**	**RR**	**P%**	**R%**
Case(1)	264	167	134	80.2	50.8
Case(2)	264	195	158	81.0	59.8

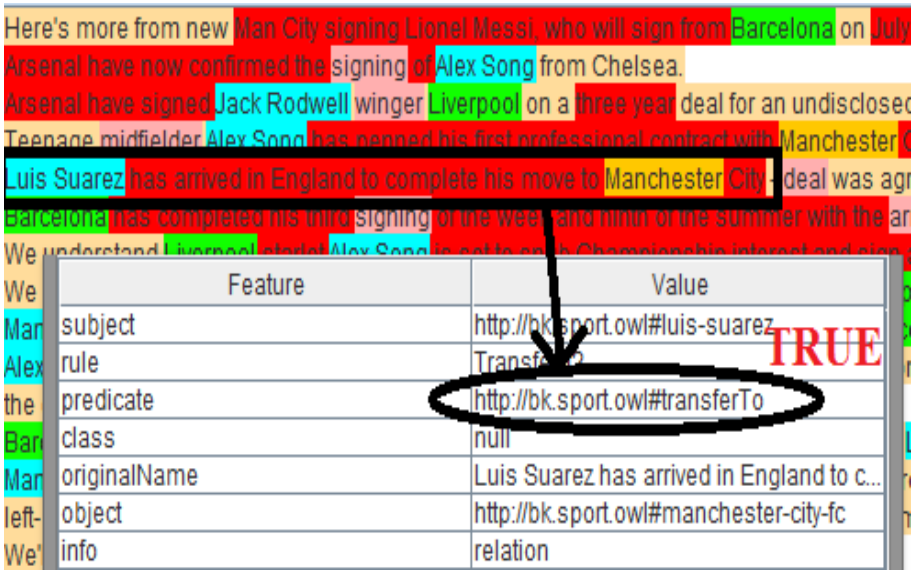
Điều này có thể được minh họa trong đoạn tin tức *"Torino have signed Serbian goalkeeper Vlada Avromov following his release from Cagliari. The 35-year-old was a free agent after leaving the Sardinian club"*. Có thể thấy rằng tin tức đó có hai quan hệ ngữ nghĩa chuyển nhượng. Đầu tiên là quan hệ ngữ nghĩa "signWith" giữa câu lạc bộ Torino với thủ môn Vlada Avramov. Thứ hai là quan hệ ngữ nghĩa "leave" (goalkeeper Vlada Avramov leaves Sardinian club). Tuy nhiên, trong trường hợp không dùng các luật nhận dạng đại từ, thì hệ thống chỉ xác định được quan hệ ngữ nghĩa đầu tiên vì trong đoạn tin tức này cụm từ "The 35-year-old" được dùng để thay thế cho "goalkeeper Vlada Avramov".

Tuy nhiên, thuật toán ở thời điểm này vẫn còn một số hạn chế. Hình 2.25 cho thấy một vài bộ ba ngữ nghĩa không nhận dạng được do cấu trúc phức tạp, đó là những quan hệ tương đương mang nhiều nghĩa nhập nhằng. Ví dụ, "Queens Park Rangers boss Harry Rednapp is eyeing a reunion with former Tottenham star Rafael van der Vaart".

Một vài trường hợp bị nhận dạng nhầm vì các lý do sau đây. Trong câu, cùng lúc có một số thực thể có tên giống nhau và hệ thống không thể nhận dạng ra được thực thể chính của quan hệ. Thông tin ngữ cảnh (mô tả cái đã không xảy ra và các sự kiện phủ định) không được bao gồm trong các từ khóa mà lại nằm trong ý nghĩa của câu. Ví dụ, một thông báo sau: "The odds on Antoine Griezmann joining Monaco have shortened again", hệ thống xác định như là <Antoine Griezmann> <transferTo> <Monaco>. Nhưng thực tế sự kiện đã không xảy ra, điều này được thể hiện trong hình 2.24. Hình 2.23 minh họa trường hợp các chú thích ngữ nghĩa được nhận dạng đúng.

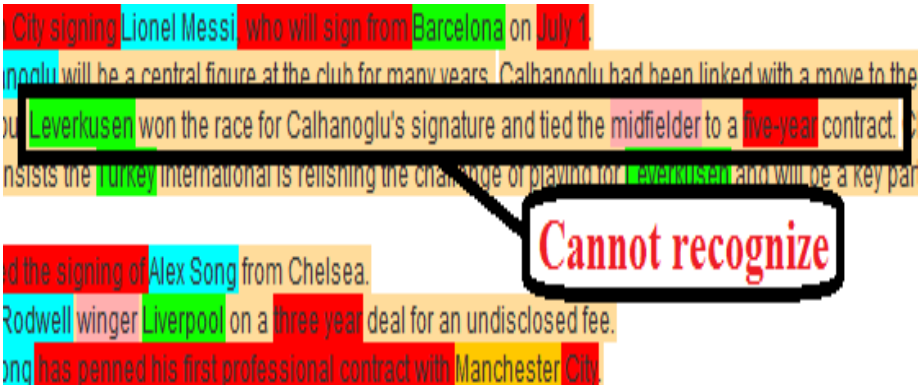
Khi phân tích kết quả thực nghiệm đầu tiên, tác giả nhận thấy độ bao phủ không cao là do cấu trúc phức tạp của các câu và chất lượng của tập từ vựng được dùng để nhận dạng động từ câu. Do đó, luận án đã thực hiện những cải tiến nhỏ bằng cách xem xét lại các quan hệ trong ontology, thêm vào từ đồng nghĩa và các biến thể của động từ vào trong bộ từ vựng. Thêm vào đó, bước tiền xử lý câu đã và đang được thực hiện để chuyển những trường hợp chắc chắn sang

dạng chuẩn. Ví dụ, <Named Entity>'s signature được chuyển thành the signature of <Named Entity>. Nhờ có bước này, các ngữ nghĩa về chuyển nhượng bóng đá được nắm bắt nhiều hơn bởi các luật nhận dạng đang có.



Hình 2.23 *Ví dụ về các chú thích nhận dạng đúng*

Hình 2.24 *Ví dụ về các chú thích nhận dạng không đúng*



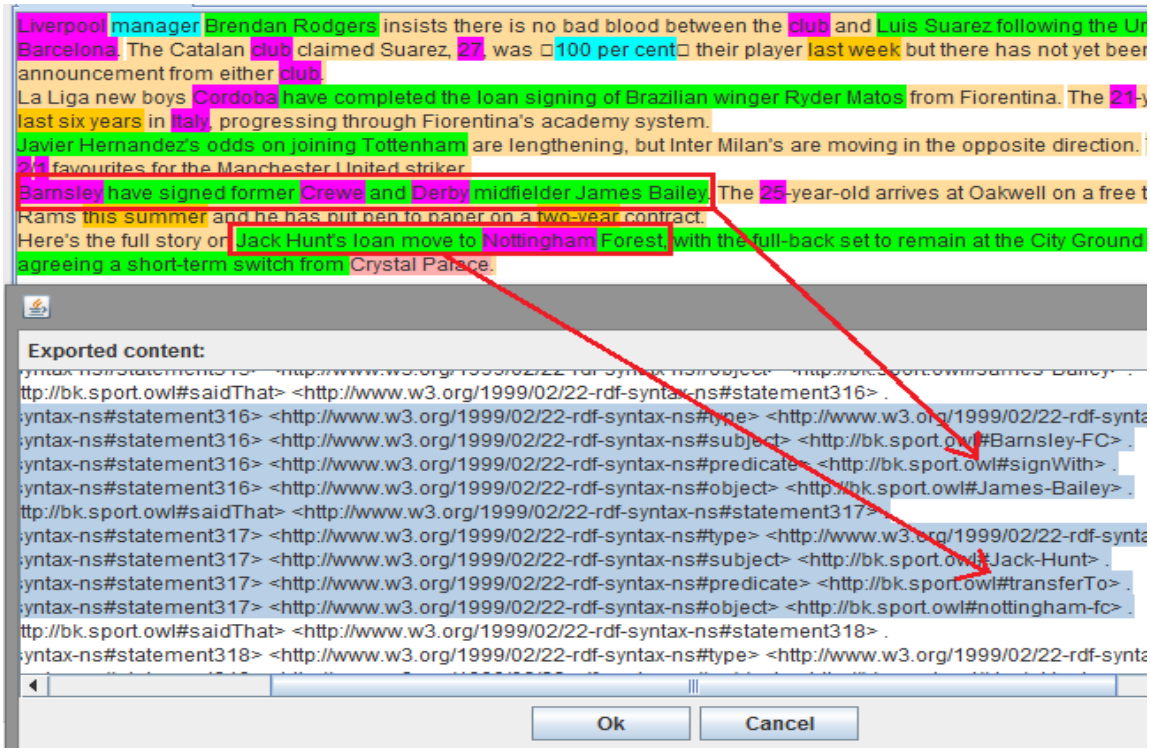
Hình 2.25 *Ví dụ về các chú thích không được nhận dạng*

Bảng 2.6 cho thấy những kết quả thực nghiệm thu được từ những nỗ lực nêu trên. Độ bao phủ được cải thiện khoảng 10% trong khi độ chính xác không thay đổi nhiều.

Bảng 2.6. *Cải thiện hiệu năng của nhận dạng quan hệ ngữ nghĩa*

	TRE	TR	RR	P%	R%
Case (1)	264	180	145	80.5	54.9
Case (2)	264	213	173	81.2	65.5

Hình 2.26 minh họa chú thích về ngữ nghĩa chuyển nhượng được sinh ra với nghiên cứu trong luận án. Các bộ ba ngữ nghĩa sau khi được trích rút được đưa vào dạng N-triples và rất dễ dàng khi chuyển chúng sang hình thức khác như RDF hay OWL.



Hình 2.26 *Các bộ ba ngữ nghĩa được trích rút là kết quả đầu ra*

2.4.3 Đánh giá chung

Những kết quả thực nghiệm cho thấy, phương pháp đề xuất trong luận án đã đáp ứng mục tiêu nghiên cứu với kết quả tích cực. Độ chính xác trong nhận dạng thực thể có tên là trên 90% và độ chính xác trong phát hiện và tạo ra chú thích ngữ nghĩa là trên 80%.

Theo hiểu biết của tác giả, nghiên cứu thực hiện trong luận án là một trong những nghiên cứu đầu tiên về sinh chú thích ngữ nghĩa về tin tức thể thao. Có nhiều nghiên cứu khác nhau về bài toán tạo ra chú thích ngữ nghĩa cho văn bản nói chung. Nhưng xét cụ thể về các dạng chú thích ngữ nghĩa mà luận án hướng tới, chưa có nghiên cứu nào đề cập đến. Đó là lý do trong các thực nghiệm, luận án chưa so sánh hiệu quả của phương pháp đề xuất với các phương pháp khác, do mục tiêu hướng đến khác nhau. Nếu áp dụng các phương pháp chú thích ngữ nghĩa tự động được đề xuất cho lĩnh vực tổng quát, vào một lĩnh vực cụ thể như thể thao thì kết quả sẽ khiêm tốn, do khả năng nhận dạng thực thể có tên hạn chế hơn. Nền tảng KIM [36] hay ASKNET [115] sẽ gặp khó khăn lớn khi nhận dạng thực thể có tên của bất kỳ cầu thủ bóng đá nổi tiếng nào trong ngữ cảnh chuyên môn của họ, và do vậy gần như không thể sinh các chú thích một cách tự động về các hoạt động, sự kiện, nhân vật thể thao. Một lý do khác là chưa có tập dữ liệu chuẩn về chú thích ngữ nghĩa cho văn bản trong lĩnh vực thể thao được công bố để đánh giá so sánh giữa nghiên cứu của luận án với các nghiên cứu liên quan, trong khi đó trong lĩnh vực y tế và sinh học đã có.

So sánh với những nghiên cứu liên quan đến trích rút thông tin ngữ nghĩa trong lĩnh vực tổng quát và lĩnh vực cụ thể như: PANKOW [34] (độ chính xác tối đa là 69%), KIM (độ chính xác là 86%, độ bao phủ là 82%), SemTag (độ chính xác là 82%) và hệ thống Asknet [115] (độ chính xác tổng thể 79.1%), tiếp cận của [38] (độ chính xác 81.2%), phương pháp sử dụng mô hình ngữ nghĩa để trích rút các quan hệ giữa các thực thể trong y học bởi [39] (độ chính xác 74.21%), mặc dù không cùng tập dữ liệu nhưng những kết quả thu được của luận án là đáng khích lệ.

2.5 Kết luận chương

Chương này trình bày những nghiên cứu về sinh chú thích ngữ nghĩa cho tin tức thể thao, đó là một phần công việc của luận án với chủ đề "Mô hình ngữ nghĩa cho hệ thống tìm kiếm tin tức thể thao". Sau khi tìm hiểu về cơ sở lý thuyết của sinh chú thích ngữ nghĩa cho tài liệu và các phương pháp tạo chú thích ngữ nghĩa, tác giả đề xuất các thuật toán sinh chú thích ngữ nghĩa cho các tin tức thể thao (cụ thể là bóng đá) và đánh giá về những hiệu quả đạt được trong các thử nghiệm ở từng nghiên cứu.

Tiếp cận nghiên cứu cho bài toán này được triển khai trong một quá trình khá dài và liên tục. Các kết quả nghiên cứu đã trình bày được tác giả công bố trong bài báo "A novel approach for automatic extraction of semantic data about football transfer in sport news" tại tạp chí *International Journal of Pervasive Computing and Communications* (2015); và trong bài báo "Automatic Semantic Annotation of Sport News Using Knowledge Base and Extraction Patterns" tại tạp chí *Journal of Science & Technology Technical Universities* (2018). Chúng tiếp nối và kế thừa một số kết quả nghiên cứu trước đó của tác giả [119] [120] và bổ sung những đóng góp mới.

Tại xuất phát điểm của nghiên cứu, tác giả xác định được ý tưởng giải quyết vấn đề sinh chú thích ngữ nghĩa từ các tin tức thể thao là dựa trên các thực thể có tên. Nghiên cứu đề xuất được phương pháp nhận dạng thực thể có tên như là các thể hiện của ontology, đạt được mức độ chi tiết hơn về ngữ nghĩa so với khái niệm nhận dạng của KIM. Sau đó, một thuật toán phát hiện ngữ nghĩa mô tả một số thông tin quan trọng và cơ bản trong các tin tức thể thao được đề xuất nhờ sự phối hợp mô tơ trích rút KIM với cơ sở tri thức và ontology thể thao được xây dựng hoàn toàn mới.

Kế thừa và cải thiện những thuật toán đã xây dựng ở giai đoạn đầu, luận án tiếp tục tập trung cải thiện hiệu quả của tác vụ nhận dạng thực thể có tên rút gọn, thực thể cùng tên khác kiểu. Quan trọng hơn, nghiên cứu đã bổ sung thêm khả năng sinh chú thích ngữ nghĩa về tuyên bố gián tiếp vào thuật toán. Việc cải tiến trong thuật toán đã mang lại những kết quả khả quan. Sau đó, tác giả tập trung giải quyết một lĩnh vực đặc thù trong tin tức thể thao đó là trích rút các quan hệ ngữ nghĩa về chuyển nhượng bóng đá dùng mô hình ngôn ngữ. Các mô hình ngôn ngữ được xây dựng dựa trên các luật nhận dạng để nắm bắt các quan hệ ngữ nghĩa. Để cải thiện độ bao phủ, tác giả đề xuất thêm một phương pháp giải quyết đồng tham chiếu thực thể dựa vào việc nhận dạng đại từ.

Có thể nói, tiếp cận xuyên suốt trong các nghiên cứu đã trình bày là sử dụng cơ sở tri thức và ontology thể thao trong việc nhận dạng thực thể có tên, và phát hiện các khái niệm (class) và một số quan hệ đặc biệt trong tin tức. Phương pháp phát hiện các bộ ba ngữ nghĩa dựa trên các luật được định nghĩa dựa trên ontology. Kết quả là chất lượng các chú thích ngữ nghĩa được phát hiện được cải thiện qua các nghiên cứu và được lưu trữ như một thành phần quan trọng của hệ thống BKSport.

Những nghiên cứu trong tương lai sẽ tập trung vào vấn đề học các luật trích rút để nâng cao khả năng mở rộng của tiếp cận. Tác giả cũng có ý định trích rút nhiều ngữ nghĩa phức tạp hơn từ các tin tức và biểu diễn chúng trong một mô hình thích hợp chẳng hạn như bộ bốn (quadruple).

CHƯƠNG 3. MỘT PHƯƠNG PHÁP TRUY VẤN TIN TỨC THỂ THAO VỚI NGÔN NGỮ TỰ NHIÊN

Chương này nghiên cứu bài toán thứ hai của luận án là chuyển đổi câu hỏi bằng ngôn ngữ tự nhiên sang câu truy vấn viết bằng cú pháp SPARQL. Sau mục các nghiên cứu liên quan, luận án trình bày các bước của phương pháp đề xuất. Đầu tiên là phân loại câu hỏi đầu vào và cấu trúc truy vấn đầu ra, tiếp theo trình bày phương pháp, thuật toán chuyển đổi câu hỏi ngôn ngữ tự nhiên sang truy vấn SPARQL. Cuối cùng là đánh giá hiệu quả của phương pháp đề xuất trên tập câu hỏi bóng đá và kết luận.

3.1 Giới thiệu

Web hiện đang là một trong những nguồn cung cấp thông tin phổ biến nhất, phục vụ một cách đầy đủ và nhanh chóng tin tức về các sự kiện diễn ra hàng ngày trên thế giới. Trong đó, tin tức về lĩnh vực thể thao thu hút được sự quan tâm của hàng triệu người đọc. Các bài viết về lĩnh vực thể thao được cập nhật liên tục trên các trang tin của Web từ rất nhiều nguồn khác nhau. Điều này dẫn đến tình trạng người đọc đối mặt với lượng thông tin rất lớn trong đó có nhiều thông tin trùng lặp, dư thừa hoặc không nằm trong sự quan tâm của họ. Chính vì vậy, vấn đề tìm kiếm thông tin một cách nhanh chóng, chính xác và tiện lợi cho người đọc luôn là một thách thức với các hệ thống tin tức trên Web. Nhiều hệ thống tìm kiếm tin tức dựa vào từ khóa đã được xây dựng [121]. Tuy nhiên, cách tìm kiếm này thường chỉ hướng tới việc trả về các tin tức mà nội dung của chúng chứa các từ khóa trong câu truy vấn chứ không phải tin tức có nội dung phù hợp với ý nghĩa của câu truy vấn. Ví dụ, nếu một người dùng muốn tìm kiếm những thông tin về việc Arda Turan chuyển đến câu lạc bộ Barcelona, anh ta/chị ta có thể sẽ sử dụng một máy tìm kiếm từ khóa truyền thống và nhập: Arda Turan transfer to Barcelona. Máy tìm kiếm sẽ trả về tất cả những tài liệu chứa một trong các từ khóa "Arda Turan", "transfer", "Barcelona". Tuy nhiên, một tài liệu chứa cả ba từ khóa này chưa hẳn đã nói về nội dung mà người dùng tìm kiếm mong muốn, do thông thường ba từ khóa này không đi liền với nhau nên không mang ý nghĩa rằng Arda Turan chuyển đến Barcelona. Việc áp dụng tìm kiếm ngữ nghĩa sẽ giúp giải quyết vấn đề trên, cho phép trả về kết quả là các tài liệu chứa chính xác nội dung theo kỳ vọng của người dùng. Hơn nữa, ngày nay bên cạnh việc đọc tin, người đọc còn có thể quan tâm đến những thông tin liên quan đến một số thực thể xuất hiện trong tin tức như những nhân vật, tổ chức, địa điểm nào đó. Đặc điểm này thể hiện rõ rệt với các tin tức thể thao, ví dụ người đọc có xu hướng theo dõi các thông tin về Lionel Messi hay Cristiano Ronaldo cùng với việc đọc các tin tức về trận siêu kinh điển (El Clásico). Vì vậy, các kênh cung cấp thông tin lớn (như BBC) đã bắt kịp nhu cầu này và tạo ra một xu hướng đổi mới trong các giao diện của hệ thống tin tức cho phép hiển thị tin tức kèm theo các thông tin nói trên. Trên cơ sở đó, luận án xác định chức năng tìm kiếm của hệ thống tin tức BKSport phải đáp ứng được cả hai yêu cầu: (1) tìm tin tức liên quan đến câu truy vấn một cách chính xác, và (2) bổ trợ thông tin kết quả của câu truy vấn trong những trường hợp có thể. Để làm tốt được điều này, hệ thống cần phải hiểu được ý nghĩa của tin tức cũng như của câu truy vấn, và phải kết hợp được tin tức và kho tri thức về miền lĩnh vực.

Trong các nghiên cứu đã trình bày ở chương 1, luận án đã đề xuất xây dựng một hệ thống tin tức thể thao dựa trên ngữ nghĩa và biện luận về những lợi ích mà hệ thống này mang lại. Để minh chứng điều đó, nghiên cứu được khởi đầu bằng việc giải quyết vấn đề tự động/bán tự động sinh chú thích ngữ nghĩa cho tin tức, biểu diễn một tin bằng cấu trúc RDF như vừa trình bày trong chương 2. Nhiệm vụ nghiên cứu tiếp theo của luận án là làm sao có thể xây dựng được một hệ thống tìm kiếm ngữ nghĩa tin tức có khả năng vừa trả về tin tức xác đáng và phù hợp với yêu cầu của người đọc, đồng thời cung cấp thêm các thông tin bổ trợ hữu ích. Sự xác đáng của tin tức cung cấp mà luận án hướng đến không nằm ở sự trùng khớp của nội dung tin tức với các từ khóa trong câu truy vấn như các chức năng tìm kiếm phổ biến trên các trang tin hiện nay. Nó phải dựa trên sự phù hợp về ý nghĩa giữa nội dung tin tức và nội dung câu truy vấn. Hơn

nữa, hệ thống tìm kiếm này phải thân thiện với người dùng là những người đọc thông thường với vốn hiểu biết tối thiểu về công nghệ.

Đã có nhiều công trình nghiên cứu về vấn đề truy hồi thông tin từ kho dữ liệu ngữ nghĩa. Trong đó, có những nghiên cứu sử dụng trực tiếp các câu lệnh SPARQL để truy vấn ra thông tin từ kho tri thức ngữ nghĩa [122]. Tuy nhiên, việc sử dụng cú pháp SPARQL có nhiều điểm hạn chế như: cú pháp ngôn ngữ truy vấn phức tạp, mặt khác lại yêu cầu người dùng phải hiểu kiến trúc bên trong của kho tri thức ngữ nghĩa. Một số nghiên cứu khác nâng cao tính thân thiện người dùng với việc cung cấp giao diện đồ họa người dùng dựa trên ontoloy để cấu trúc nên (formulate) câu truy vấn SPARQL [123]. Tuy nhiên, các nghiên cứu trên vẫn đòi hỏi người dùng thực hiện một số thao tác nhất định và phải hiểu biết cơ bản về ontology. Vì vậy, mục tiêu của luận án là xây dựng một hệ thống tìm kiếm bằng ngôn ngữ tự nhiên, thân thiện với người dùng, không đòi hỏi họ phải có kiến thức về ngôn ngữ truy vấn phức tạp mà vẫn có thể sử dụng hiệu quả hệ thống. Thành phần tìm kiếm tin tức trong hệ thống tin tức thể thao mà luận án hướng tới bao gồm 2 thành phần: (1) thành phần đầu tiên đảm nhận việc chuyển đổi câu truy vấn ở dạng ngôn ngữ tự nhiên về dạng truy vấn có cấu trúc SPARQL, và (2) thành phần thứ hai nhận nhiệm vụ sử dụng câu truy vấn SPARQL thu được để truy vấn vào kho dữ liệu ngữ nghĩa và trả về tin tức liên quan kết hợp với câu trả lời cho câu truy vấn. Do chức năng của thành phần thứ hai đã được cung cấp bởi các mô tơ tìm kiếm ngữ nghĩa như Allegrograph, nên việc xây dựng chúng nằm ngoài phạm vi nghiên cứu của luận án. Tóm lại, chương này trình bày nghiên cứu về một phương pháp chuyển đổi câu hỏi về tin tức dưới dạng ngôn ngữ tự nhiên sang truy vấn ngữ nghĩa dạng SPARQL. Chương này tập trung vào trình bày các phương pháp và kỹ thuật để xây dựng thành phần thứ nhất (1).

Các mục còn lại của chương 3 được tổ chức như sau: mục 3.2 điểm qua một số hướng nghiên cứu liên quan đến việc xây dựng hệ thống tìm kiếm truy hồi thông tin nhất là những hệ thống lớn cho phép người dùng sử dụng các câu truy vấn dạng ngôn ngữ tự nhiên để tìm kiếm thông tin từ kho dữ liệu ngữ nghĩa. Mục 3.3 trình bày về phân loại các dạng câu truy vấn của người đọc tin tức và giới thiệu về mô hình biểu diễn ngữ nghĩa của tin tức thể thao cùng các truy vấn ngữ nghĩa – chính là đầu ra của hệ thống. Mục 3.4 trình bày phương pháp chuyển đổi từ câu truy vấn ngôn ngữ tự nhiên sang truy vấn ngữ nghĩa SPARQL, bao gồm cả quá trình xử lý chung và nguyên lý hoạt động chi tiết của từng thành phần (pha) trong hệ thống hỏi đáp. Mục 3.5 trình bày đánh giá và nhận xét kết quả thử nghiệm. Kết luận và đề xuất hướng cải tiến trong những nghiên cứu trong tương lai là nội dung của mục 3.6.

3.2 Các nghiên cứu liên quan

Ở một khía cạnh nào đó có thể nói các hệ thống QA (Question Answering) được sinh ra từ những nghiên cứu về truy hồi thông tin từ các kho dữ liệu, thông tin lớn. Sự phát triển của Web ngữ nghĩa đã mở ra một hướng đi mới trong nghiên cứu về lĩnh vực này.

Trong nghiên cứu của [124], tác giả cho rằng việc khai thác được tri thức có ý nghĩa quan trọng trong việc cải thiện được tính hiệu quả của hệ thống hỏi đáp, và các kỹ thuật Web ngữ nghĩa hỗ trợ tốt việc này. Tìm kiếm ngữ nghĩa với những thế mạnh vốn có ngay lập tức đã được quan tâm trong các nghiên cứu đầu tiên. Ban đầu, các nghiên cứu còn xuất phát từ ý tưởng đơn giản chỉ là làm sao có thể thực hiện được các truy vấn ngữ nghĩa trong các hệ thống thông tin. Trong [122], các tác giả xây dựng một ontology về thể thao với mục đích sử dụng nó để truy hồi ngữ nghĩa thông tin thể thao trên World Wide Web. Việc tìm kiếm ngữ nghĩa được thực hiện bằng cách gửi trực tiếp các câu truy vấn SPARQL vào hệ thống. Một số nghiên cứu tập trung vào việc tạo ra truy vấn SPARQL từ các giao diện đồ họa người dùng được xây dựng dựa trên ontology [123] [93].

Một số công trình dựa trên ngôn ngữ tự nhiên có kiểm soát, như Squall2Sparql [42] và GiNSENG [43], thường xem xét một tập hợp con hạn chế và rõ ràng của ngôn ngữ tự nhiên mà có thể được dịch trực tiếp sang SPARQL. Mặc dù đem lại độ chính xác cao, cách tiếp cận này có hạn chế ở sự linh hoạt và khó có khả năng áp dụng ở một lĩnh vực khác.

Đã có một số nỗ lực trong việc cải thiện tương tác trong các hệ QA (Question Answering) để hướng tới việc hỏi đáp bằng ngôn ngữ tự nhiên. Nghiên cứu của [125] đã nêu lên tầm quan

trọng và tính khả thi của một hệ thống hỏi đáp bằng ngôn ngữ tự nhiên Trung Quốc. Hệ thống hỏi đáp của họ được xây dựng dựa trên ba mô hình: mô hình hiểu ngữ nghĩa của câu hỏi dựa trên ontology và Web ngữ nghĩa, mô hình so khớp độ tương tự câu hỏi dựa trên FAQ (Frenquently Asked Questions), mô hình tự động tìm nạp câu trả lời dựa trên kho lưu trữ văn bản. Nó được cài đặt với 2 mô đun chính: mô đun phân tích câu hỏi và mô đun trích xuất câu trả lời. Với một câu truy vấn đầu vào, mô đun phân tích câu hỏi sẽ sinh ra một vài chuỗi viết lại có trọng số, sau đó chuyển truy vấn thành một véc tơ. Đồng thời, ở mô đun này còn có một bộ phân loại câu hỏi, nhằm xác định kiểu của câu trả lời cần trả về. Mô đun trích xuất câu trả lời bao gồm thành phần truy hồi tài liệu, thành phần tìm kiếm đoạn văn và so khớp câu trả lời. Cuối cùng, hệ thống sẽ tính trọng số cho các câu trả lời và đưa ra câu trả lời có trọng số lớn nhất.

Trong vài năm gần đây, một số hệ thống hỏi đáp bằng ngôn ngữ tự nhiên tiếng Anh cũng được phát triển. Điều này cho thấy nhu cầu cần được cung cấp thông tin từ người dùng đối với các nguồn tin và nguồn tri thức là rất lớn.

PANTO [126] là một giao diện ngôn ngữ tự nhiên khả chuyển tới các ontology cho phép người dùng biểu diễn nhu cầu thông tin của mình bằng ngôn ngữ tự nhiên mà không cần quan tâm đến cú pháp RDF hay OWL, ngôn ngữ truy vấn SPARQL và từ vựng của ontology. Nó sử dụng từ điển tổng hợp Wordnet và thuật toán đo chuỗi để ánh xạ các từ trong câu truy vấn người dùng vào các thành phần trong ontology (khái niệm, thể hiện, quan hệ). Nó sử dụng bộ phân tích cú pháp StandfordParser để phân tích câu hỏi đầu vào thành cây phân tích, sau đó trích xuất các cụm từ danh định để hình thành nên dạng biểu diễn trung gian QueryTriples. Để truy vấn ontology, biểu diễn trung gian này sau đó sẽ được ánh xạ sang dạng OntoTriples. Hệ thống thử nghiệm trên bộ dữ liệu được cung cấp bởi Mooney và đạt độ chính xác tốt nhất là 90.87% và độ bao phủ tốt nhất là 96.64% cho bộ dữ liệu về nhà hàng. Tuy nhiên, hệ thống vẫn còn hạn chế trong việc xử lý các câu hỏi phủ định và chưa xử lý được các câu hỏi về số lượng.

Querix [127] là một hệ thống hỏi đáp được hỗ trợ ontology, dựa trên việc yêu cầu người dùng làm rõ các trường hợp nhập nhằng bằng cách hiện ra các hộp thoại cho người dùng lựa chọn. Hệ thống này gồm các thành phần giao diện người dùng, bộ quản lý ontology, bộ phân tích truy vấn, trung tâm so khớp, bộ sinh truy vấn, thành phần hộp thoại, và lớp truy cập ontology. Querix sử dụng từ điển tổng hợp Wordnet để xác định các từ đồng nghĩa cho các từ trong câu hỏi ngôn ngữ tự nhiên đầu vào nhằm mục đích nhận diện các thể hiện xuất hiện trong câu hỏi được đầy đủ hơn. Việc xác định các quan hệ bộ ba của Querix dựa trên việc ánh xạ dãy các loại từ chính (kết quả của bước phân tích cú pháp bằng Stanford Parser) với một tập các mẫu horixtic. Thử nghiệm trên ontology được xây dựng dựa vào cơ sở tri thức thông tin địa lý về nước Mỹ của Mooney và các đồng sự. Hệ thống được chạy thử nghiệm trên 215 câu hỏi khác nhau đạt độ chính xác là 77.67% và độ bao phủ là 78.6%.

QuestIO (Question-based Interface to Ontologies) [128] là một công cụ phục vụ cho việc truy vấn kho tri thức lớn lưu trữ trong ontology sử dụng ngôn ngữ tự nhiên. Điểm đặc biệt của công cụ là nó độc lập về miền. Chính vì điều này mà QuestIO có thể được nhúng vào bất kỳ hệ thống nào hoặc được sử dụng với bất kỳ ontology hoặc cơ sở tri thức nào mà không cần phải tùy chỉnh. Điểm hạn chế của công cụ này đến từ việc nhận dạng quan hệ xuất hiện trong câu truy vấn đầu vào dựa trên luật mà không phân tích cú pháp câu truy vấn ở mức sâu, do đó không xử lý được những câu truy vấn có ngữ nghĩa phức tạp. Tiến hành thử nghiệm trên bộ dữ liệu gồm 22 câu hỏi từ danh sách gửi thư người dùng GATE (đây là nơi mà người dùng hỏi về các mô đun và các plugin đa dạng của GATE), công cụ đạt độ chính xác 71.88%.

FREyA [129] là phiên bản phát triển của QuestIO. Điểm vượt trội của FREyA so với phiên bản trước đó là thay vì dùng luật để phát hiện quan hệ có trong câu hỏi tự nhiên đầu vào, FREyA sử dụng phương pháp kết hợp phân tích cú pháp với tìm kiếm dựa trên ontology. Điều này khiến FREyA có khả năng xử lý được những câu hỏi có ngữ nghĩa phức tạp hơn. Hơn nữa, trong trường hợp hệ thống không tự động lấy được một câu trả lời, nó sẽ hiển thị ra hộp thoại để người dùng lựa chọn. Sự lựa chọn của người dùng sau đó sẽ được lưu lại để cải thiện hệ thống. Tiến hành thử nghiệm trên bộ dữ liệu Mooney Geoquery gồm 250 câu hỏi, hệ thống đạt độ chính xác và độ truy hồi bằng nhau và bằng 92.4%.

ORAKEL [130] đem đến một giao diện ngôn ngữ tự nhiên có khả năng chuyển đổi câu truy vấn ngôn ngữ tự nhiên về dạng câu truy vấn có cấu trúc ứng với một ontology cho trước. Sự chuyển đổi này được thực hiện dựa vào bộ diễn dịch truy vấn (diễn dịch câu hỏi đầu vào và chuyển nó về biểu diễn dưới dạng logic bậc nhất) và bộ chuyển đổi truy vấn (chuyển từ dạng biểu diễn logic của câu hỏi về dạng truy vấn SPARQL). Nó yêu cầu hai bộ từ vựng: bộ từ vựng về một miền cụ thể và bộ từ vựng độc lập về miền. Hệ thống này chỉ cần một ontology cho trước và một bộ từ vựng về một miền cụ thể là có thể hoạt động được. Hạn chế của hệ thống là chỉ xử lý được những câu hỏi có từ để hỏi (dạng wh-question), trong khi không xử lý được những câu hỏi không có từ hỏi (dạng yes/no-question).

PowerAqua [131] là một hệ thống hỏi đáp đa ontology, nhận đầu vào là một câu truy vấn dạng ngôn ngữ tự nhiên và trả về câu trả lời lấy ra từ các nguồn khác nhau trên Web ngữ nghĩa. Điểm đặc biệt của PowerAqua là nó không thiết kế để hướng tới một ontology cụ thể cho miền nào cả, vì thế nó là hệ thống mạnh về hỏi đáp trên miền dữ liệu lớn và không đồng nhất. Cách tiếp cận của PowerAqua là ánh xạ câu truy vấn dạng ngôn ngữ tự nhiên về một biểu diễn bộ ba. Sau đó, bằng việc sử dụng các tìm kiếm horixtic, nó sẽ trả về các đồ thị con phù hợp từ kho ngữ liệu RDF. Tuy nhiên, PowerAqua chỉ xử lý tốt những câu truy vấn có cấu trúc đơn giản, dễ dàng chuyển đổi về dạng biểu diễn bộ ba. Nó sẽ xử lý sai nếu câu truy vấn của người dùng phức tạp hơn, ví dụ như những câu hỏi chứa "the most", "at least" hoặc "more than", "less than". Trong khi nếu áp dụng việc phân tích sâu hơn về cấu trúc ngữ pháp của câu truy vấn đầu vào, các câu có dạng như trên có thể được xử lý chính xác.

AquaLog [132] là một hệ thống hỏi đáp khả chuyển, nhận một câu truy vấn ở dạng ngôn ngữ tự nhiên và một ontology làm đầu vào, trả về các câu trả lời lấy từ một hoặc nhiều cơ sở tri thức. AquaLog kết hợp sử dụng nền tảng xử lý ngôn ngữ tự nhiên GATE, các thuật toán đo khoảng cách chuỗi ký tự, từ điển tổng hợp WordNet, và một dịch vụ tính độ tương đồng dựa trên ontology cho các quan hệ và các lớp để ánh xạ các thành phần trong câu truy vấn đầu vào đến ontology và cơ sở tri thức mục tiêu. AquaLog còn áp dụng kỹ thuật học máy để trích rút quan hệ giữa các đối tượng, tuy nhiên chỉ thực hiện một cách bán tự động. Điểm hạn chế của AquaLog cũng tương tự như PowerAqua đến từ cơ chế so khớp cấu trúc cú pháp của câu truy vấn với một số mẫu cấu trúc có sẵn. Do đó phạm vi hoạt động hiệu quả của nó bị giới hạn, nó chỉ xử lý tốt đối với những câu có cấu trúc đơn giản.

Pythia [133] cũng là một hệ thống hỏi đáp nhận đầu vào là câu truy vấn ở dạng ngôn ngữ tự nhiên. Tuy nhiên, cách tiếp cận để xử lý câu truy vấn đầu vào của nó khác với hai hệ thống PowerAqua và AquaLog. Pythia phân tích cấu trúc ngữ pháp của câu truy vấn đầu vào một cách sâu hơn. Vì thế, nó có thể xử lý được những câu truy vấn đầu vào phức tạp, như các câu chứa cụm từ "more than", "the most". Tuy nhiên, điểm yếu của Pythia là nó hoạt động dựa trên bộ từ điển các biểu diễn ngữ nghĩa của một ontology cho trước. Bộ từ điển này được xây dựng một cách thủ công, vì thế nó sẽ không khả thi khi triển khai trong các tập dữ liệu kích thước rất lớn.

Trong bài báo “Template-based Question Answering over RDF Data” [134], Unger và các đồng nghiệp trình bày một cách tiếp cận cũng dựa trên phân tích cấu trúc ngữ pháp của câu truy vấn đầu vào. Ý tưởng đề xuất là tiến hành phân tích cấu trúc cây của câu truy vấn để sinh ra một mẫu truy vấn SPARQL. Bước này bao gồm bước con nhận dạng thực thể thống kê và bước con phát hiện vị ngữ. Mặc dù hệ thống có thể xử lý được những câu truy vấn có cấu trúc phức tạp như hỏi về số lượng, so sánh hơn, so sánh hơn nhất, nó vẫn chưa xử lý được những dạng câu có nhiều chủ thể, nhiều tân ngữ hay câu có đề cập đến ngữ cảnh thời gian.

Các nghiên cứu trên cho thấy ý nghĩa của việc thực hiện tính năng tìm kiếm ngữ nghĩa thông qua hình thức truy vấn diễn đạt bằng ngôn ngữ tự nhiên. Tuy nhiên chúng tập trung vào miền ứng dụng chung hơn là các miền ứng dụng đặc thù. Khi ứng dụng vào lĩnh vực đặc biệt như thể thao, với những đặc tính riêng - kết quả tìm kiếm thu được là chưa chính xác. Các dạng câu truy vấn mà các hệ thống trên nhận biết được cũng có cấu trúc thường là đơn giản, chưa diễn tả được hết nhu cầu thông tin của độc giả. Trong bối cảnh phát triển hệ thống BKSport, vấn đề chính cần quan tâm là tìm kiếm tin tức thể thao liên quan tới nhu cầu độc giả rồi sau đó mới tới hiển thị thông tin bổ trợ. Do đó, cần có tiếp cận riêng để cải thiện hơn nữa kết quả tìm kiếm. Để làm điều đó, việc chuyển đổi truy vấn sang dạng có ngữ nghĩa cần được nâng cao độ chính

xác. Tác giả đề xuất một phương pháp mới gồm nhiều giai đoạn nhằm thực hiện điều này trong lĩnh vực tin tức thể thao. Trong đó việc mô hình hóa câu hỏi, phân tích nhận biết cấu trúc ngữ pháp và chuyển đổi sang dạng biểu diễn ngữ nghĩa tương ứng đóng vai trò quyết định.

3.3 Phân loại câu hỏi đầu vào và cấu trúc truy vấn đầu ra

3.3.1 Phân loại câu hỏi

Trước khi nhận biết được câu truy vấn và chuyển đổi chúng sang dạng ngữ nghĩa, ta cần phân loại chúng. Có nhiều yếu tố quyết định việc phân loại. Dựa vào cấu trúc ngữ pháp, luận án phân chia câu truy vấn thành hai loại: câu hỏi có từ hỏi và câu hỏi nghi vấn (yes/no). Đối với dạng câu hỏi có từ hỏi, dựa vào những thông tin thể thao mà người dùng quan tâm, luận án tập trung xử lý các dạng câu hỏi có từ hỏi như who, which, what, where và how many.

Ngoài ra, câu truy vấn còn có thể được phân loại dựa trên loại của câu trả lời kỳ vọng và được mô tả trong hình 3.1. Từ ý tưởng này, luận án phân loại các câu truy vấn thành các loại như sau:

Câu hỏi vị ngữ (Predicative question), ví dụ:

- Which team defeated Chelsea this season?
- Who transferred to Barcelona this year?
- Which news is about Lionel Messi?
- Whom did Wayne Rooney dispute with?

Câu hỏi nghi vấn (Yes/No question), ví dụ:

- Did Real Madrid win Bayern Munich yesterday?
- Was Barcelona defeated by Chelsea yesterday?

Câu hỏi về định nghĩa (Definition question), ví dụ:

- Who is Lionel Messi?
- What is FIFA?

Câu hỏi kết hợp (Association question), ví dụ:

- What is result of the match between Real Madrid and Barcelona?
- What happen between Real Madrid and Barcelona?

Câu hỏi số lượng (Quantity question), ví dụ:

- How many goals were scored by Lionel Messi yesterday?

Câu hỏi nhiều chủ ngữ, nhiều tân ngữ (Multi-subject, multi-object question), ví dụ:

- Which team defeated Chelsea and Barcelona?
- Did Chelsea and Real Madrid defeat Barcelona in this season?

Câu hỏi về ý kiến (Opinion question), ví dụ:

- What did Lionel Messi say/think/about Manchester United?

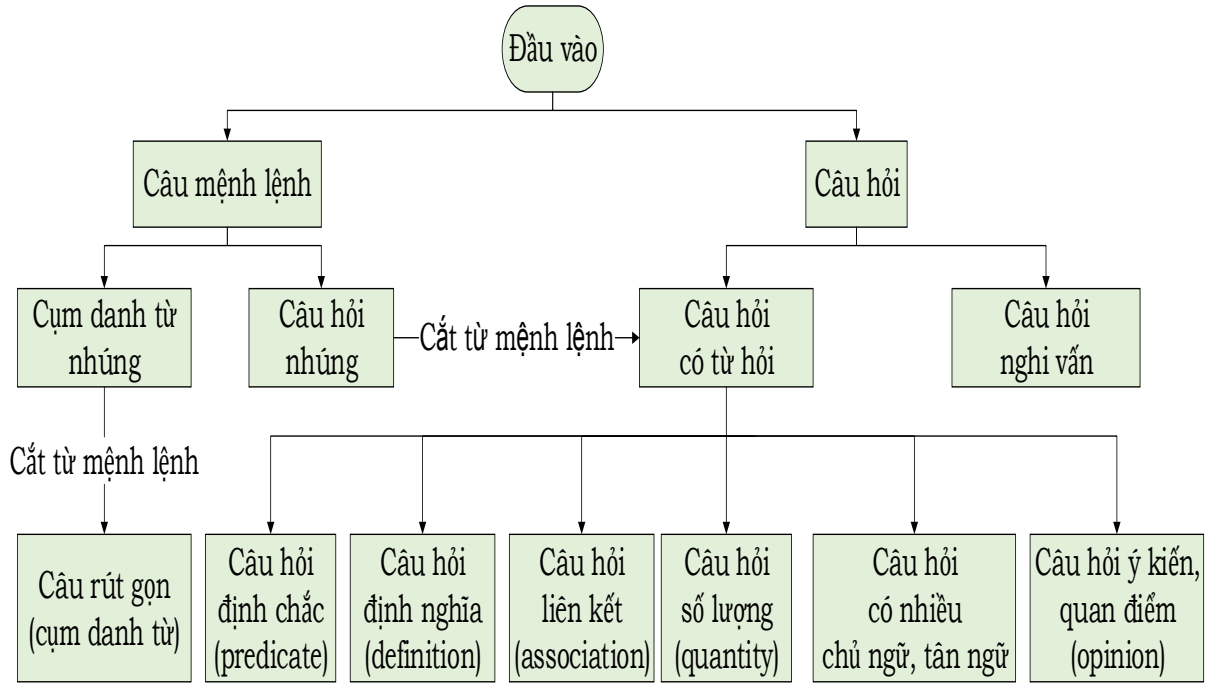
Câu hỏi so sánh hơn, so sánh hơn nhất (Comparative, superlative question), ví dụ:

- Which team won 3 games this year?
- Which news contains at least 2 teams?

Thêm vào đó, hệ thống cũng chấp nhận câu mệnh lệnh như:

Câu mệnh lệnh (Imperative sentence), ví dụ:

- Show me news about Lionel Messi.
- Give me result of the match between Chelsea and Barcelona.



Hình 3.1 *Phân loại các câu truy vấn*

Hệ thống mà luận án xây dựng không đơn thuần là một hệ thống hỏi đáp ngữ nghĩa (semantic-based QA). Nó còn được xây dựng theo tiêu chí trợ giúp tối đa người dùng đọc tin tức, giúp cho họ không chỉ tìm kiếm tin tức một cách chính xác và nhanh chóng, mà còn trả về những thông tin tri thức có liên quan. Do đó, đối với bất kỳ loại câu truy vấn nào nêu trên, hệ thống đều sẽ chuyển đổi về câu truy vấn dạng SPARQL, và đảm bảo rằng từ câu truy vấn SPARQL được sinh ra đó các tin tức và các thông tin khác liên quan sẽ được trả về nếu chúng tồn tại trong cơ sở tri thức.

3.3.2 Chú thích và truy vấn ngữ nghĩa về tin tức thể thao

Các truy vấn ngữ nghĩa là đầu ra của phương pháp đề xuất trong nghiên cứu này, vì vậy trước tiên ta cần xác định chúng sẽ được biểu diễn như thế nào. Để có thể tìm kiếm thông tin từ kho dữ liệu ngữ nghĩa, các câu truy vấn cần có cấu trúc phù hợp với cấu trúc diễn đạt của dữ liệu trong kho dữ liệu ngữ nghĩa. Tiểu mục này trình bày về phương pháp biểu diễn tin tức dưới dạng các chú thích ngữ nghĩa và các mẫu truy vấn ngữ nghĩa tương ứng.

Luận án xác định những nội dung xoay quanh một tin tức cần được chú thích ngữ nghĩa bao gồm hai kiểu:

(1) Các thuộc tính của tin tức (ví dụ như: URL, createtime, chủ đề …), các thực thể được nhắc đến trong tin tức (ví dụ như: cầu thủ, đội bóng, giải đấu …), và các thuộc tính của các thực thể này (ví dụ như: vị trí chơi của cầu thủ, đội bóng mà cầu thủ đang chơi, giải đấu mà đội bóng tham gia ...). Đối với những nội dung này, luận án biểu diễn chú thích ngữ nghĩa dưới dạng RDF.

Ví dụ:

```
<owl:NamedIndividual rdf:about="http://bk.sport.owl#jonathan-viera">
  <bksport:playFor rdf:resource="http://bk.sport.owl#ud-las-palmas"/>
  <protons:generatedBy rdf:resource="http://bk.sport.owl"/>
  <protons:hasAlias>Jonathan Viera</protons:hasAlias>
  <rdfs:label>Jonathan Viera</rdfs:label>
  <protons:mainLabel>Jonathan Viera</protons:mainLabel>
```

```
        <rdf:type rdf:resource="http://bk.sport.owl#Midfield"/>

    </owl:NamedIndividual>
```

(2) Các hoạt động mà tin tức đề cập đến (ví dụ: cầu thủ ghi bàn, chuyển nhượng cầu thủ, trận đấu giữa hai đội bóng …). Đối với những nội dung này, luận án đề xuất bộ bốn (quadruple) để chú thích ngữ nghĩa, vì đây là các sự kiện diễn ra trong ngữ cảnh của tin tức.
Ví dụ:

```
PREFIX bksport: <http://bk.sport.owl#>

<bksport#Romelu_Lukaku> <bksport#playFor> <bksport#manchester_united_fc> <bksport#namedgraph> .
```

Để truy vấn các thông tin trong cơ sở tri thức, luận án sử dụng cú pháp truy vấn SPARQL. Ngôn ngữ SPARQL là một ngôn ngữ truy vấn ngữ nghĩa cho cơ sở dữ liệu, có khả năng truy hồi và thao tác trên các dữ liệu được lưu trữ ở định dạng RDF.

Tương ứng với hai dạng biểu diễn chú thích ngữ nghĩa trên, luận án cũng định nghĩa hai dạng khác nhau của câu truy vấn ngữ nghĩa SPARQL như sau:

- Đối với những biểu diễn bộ ba, câu truy vấn SPARQL sẽ có khung như sau:

```
SELECT/ASK
WHERE
{
        // Query triple patterns
}
```

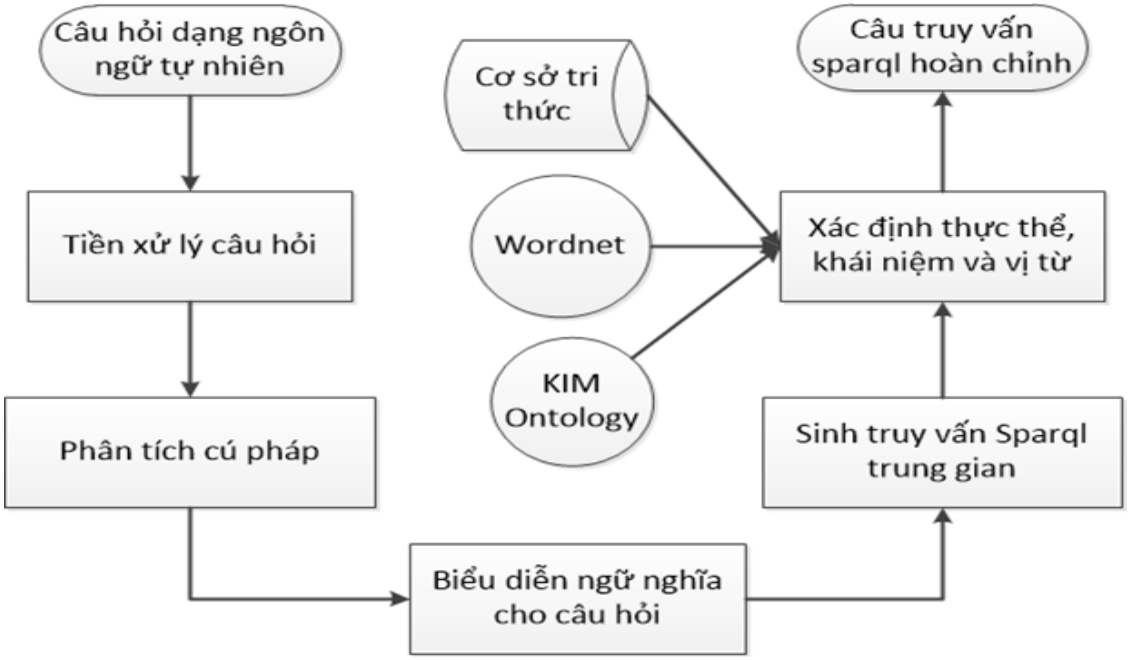
- Đối với những biểu diễn bộ bốn, câu truy vấn SPARQL sẽ có khung như sau:

```
SELECT/ASK
WHERE
{
        ?namedgraph
        {
                // Query triple patterns

        }
}
```

3.4 Phương pháp chuyển đổi câu hỏi ngôn ngữ tự nhiên sang truy vấn SPARQL

Từ mục tiêu nghiên cứu và kết quả khảo sát các nghiên cứu liên quan, tác giả đề xuất một phương pháp cho phép chuyển đổi các câu hỏi dưới dạng ngôn ngữ tự nhiên về tin tức thể thao sang các câu truy vấn biểu diễn bằng ngôn ngữ SPARQL. Phương pháp luận án đề xuất được mô tả trong hình 3.2 gồm 5 giai đoạn chính. Quy trình xử lý diễn ra cụ thể như sau. Câu hỏi đầu vào dạng ngôn ngữ tự nhiên trước tiên sẽ được mô đun tiền xử lý chuẩn hóa để các mô đun sau hoạt động hiệu quả và chính xác. Câu hỏi sau khi được tiền xử lý sẽ đi qua bộ phân tích cú pháp để phân tích các thành phần ngữ pháp và mối quan hệ giữa các thành phần ngữ pháp đó, từ đó biểu diễn câu hỏi dưới dạng mô hình ngữ nghĩa. Từ mô hình ngữ nghĩa, câu truy vấn SPARQL trung gian được sinh ra. Mô đun phát hiện thực thể có tên, khái niệm và vị từ sẽ chú thích các biến trong câu truy vấn SPARQL trung gian bằng các URI trong ontology và cơ sở tri thức của hệ thống. Cuối cùng, câu truy vấn SPARQL hoàn chỉnh được sinh ra.



Hình 3.2 *Quy trình chuyển đổi câu hỏi từ ngôn ngữ tự nhiên sang SPARQL*

3.4.1 Tiền xử lý câu hỏi

Mô đun tiền xử lý có nhiệm vụ chuẩn hóa câu hỏi đầu vào ở dạng ngôn ngữ tự nhiên để nâng cao hiệu quả xử lý cho các mô đun sau. Những công việc tiền xử lý bao gồm:

- Chuẩn hóa những token không chuẩn: khi viết người dùng thường có thói quen sử dụng nhiều ký hiệu viết tắt. Luận án thống kê những ký hiệu viết tắt thông thường và xây dựng nên bảng chuẩn hóa gồm hai trường. Trường thứ nhất là những ký hiệu viết tắt thường dùng và trường thứ hai là những từ ngữ có ý nghĩa tương đương. Sau đó, luận án tiến hành duyệt từng token trong câu truy vấn, những ký hiệu viết tắt sẽ được thay thế bằng nhóm từ tương ứng.
- Xác định thuộc tính thời gian của câu truy vấn: luận án thống kê và phân loại các nhãn thời gian thành các loại như sau:
    + Khoảng thời gian một ngày: ví dụ “today”, “yesterday”, …
    + Khoảng thời gian một tuần: ví dụ “next week”, “last week”, …
    + Khoảng thời gian một tháng: ví dụ “next month”, “last month”, …
    + Khoảng thời gian một năm: ví dụ “next year”, “last year”, …

Dựa vào thời điểm người dùng truy vấn, hệ thống tính ra một giá trị thời gian cụ thể tương ứng với từng loại nhãn thời gian, sau đó thay thế các nhãn thời gian trong câu truy vấn bằng các giá trị cụ thể này.

Chuyển đổi tương đương giữa các truy vấn: hệ thống chấp nhận cả những đầu vào là câu mệnh lệnh hoặc câu rút gọn. Để bộ phân tích cú pháp hoạt động đúng cũng như để đơn giản hóa việc xử lý ở các bước sau, hệ thống chuyển đổi những câu như vậy về một trong hai dạng câu hỏi chuẩn có ý nghĩa tương đương: dạng câu hỏi có từ hỏi hoặc dạng câu hỏi nghi vấn (yes/no). Ví dụ đối với câu hỏi: “news about Lionel Messi” sẽ được chuyển đổi thành câu hỏi chuẩn ngữ pháp hơn là “Which news is about Lionel Messi?”.

3.4.2 Phân tích cú pháp

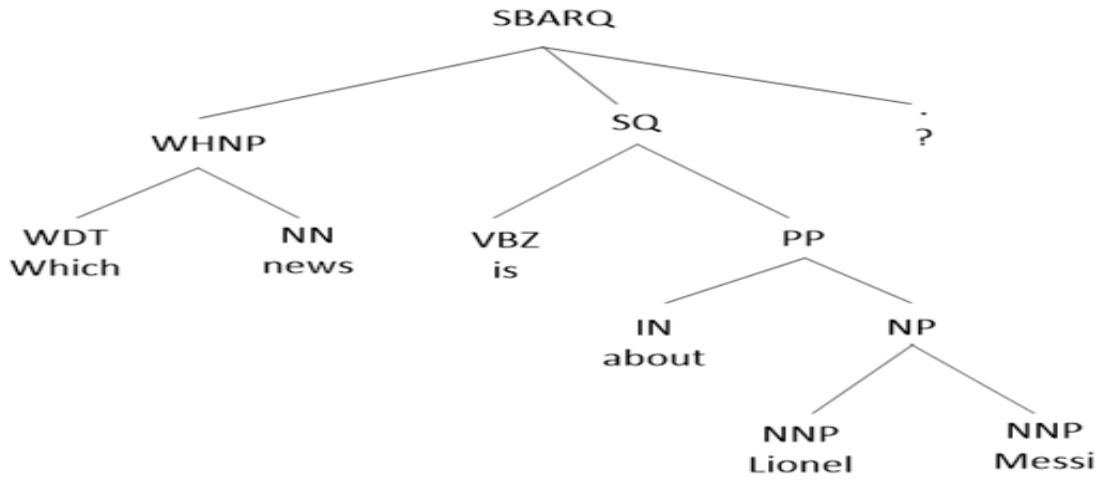
Đây là một giai đoạn quan trọng, ảnh hưởng nhiều tới kết quả cuối cùng. Việc cần làm là phải phân tích được các thành phần ngữ pháp của câu truy vấn tự nhiên, để từ đó có thể chuyển đổi chúng sang các thành phần cấu trúc của truy vấn SPARQL. Mô đun phân tích cú pháp xác định dạng thức câu truy vấn, các thành phần ngữ pháp trong câu truy vấn và mối quan hệ giữa chúng. Để làm được tất cả những điều này, tác giả tiến hành phân tích gắn nhãn từ loại (Part-Of-Speech Tagging), cây cấu trúc cụm từ (Phrase Structure Tree) và các phụ thuộc theo loại

(Typed Dependencies). Kết quả của các bước phân tích trên sẽ được sử dụng trong các tác vụ ở giai đoạn sau như xác định dạng câu truy vấn, xây dựng các quan hệ bộ ba, chú thích thực thể, lớp và thuộc tính.

Cây cấu trúc cụm từ là một cách trực quan để biểu diễn đầu ra của quá trình phân tích cú pháp của câu. Nó chỉ ra ba khía cạnh của cấu trúc câu như sau:

- Thứ tự tuyến tính của các từ trong câu.
- Các nhóm từ đi với nhau tạo thành cụm từ.
- Cấu trúc phân cấp của các cụm từ.

Ví dụ, với câu truy vấn "Which news is about Lionel Messi?", ta thu được một cây cấu trúc cụm từ được minh họa trong hình 3.3 dưới đây:



Hình 3.3 *Ví dụ về cây cấu trúc cụm từ trong câu*

Trong đó, nút gốc của cây xác định dạng của câu truy vấn.

Phụ thuộc theo loại là các biểu diễn của các quan hệ ngữ pháp giữa các từ trong một câu. Chúng dễ hiểu và hữu ích cho những ai muốn trích rút các quan hệ trong văn bản. Mỗi phụ thuộc theo loại là một bộ ba của: tên quan hệ, thành phần điều khiển, và thành phần phụ thuộc. Ví dụ, đối với câu truy vấn "Which news is about Lionel Messi?", hệ thống phân tích ra các phụ thuộc theo loại như sau:

det(news-2, Which-1)
attr(is-3, news-2)
root(ROOT-0, is-3)
nn(Messi-6, Lionel-5)
prep_about(is-3, Messi-6)

Ở đây các từ viết tắt được định nghĩa trong bài báo [135] như sau:

Det (viết tắt của determiner) là quan hệ giữa phần đầu của một cụm danh từ và từ hạn định (determiner) của nó. Một số ví dụ về từ hạn định là: a, an, the, some, which, your ...

Attr (viết tắt của attributive) là quan hệ dành cho bổ ngữ của một động từ nối (copular verb) như "to be (is, am, are, was, were)", "to seem", "to appear", "to look", "to sound", "to smell", "to taste" …

Root: quan hệ ngữ pháp gốc (root) chỉ đến gốc của câu.

Nn: là quan hệ giữa một tính từ ghép với một danh từ.

Prep: từ bổ nghĩa giới từ của một động từ, tính từ hoặc danh từ là một cụm giới từ đảm nhiệm thay đổi ý nghĩa của động từ, tính từ, danh từ hoặc thậm chí một giới từ khác.

Trong phân tích cú pháp, có nhiều nhóm phụ thuộc theo loại, nhưng trong nghiên cứu này luận án chỉ quan tâm đến một số nhóm phụ thuộc theo loại nhất định. Chúng thể hiện một số dạng phụ thuộc (ràng buộc) giữa các thành phần của câu truy vấn, ví dụ như: chủ thể - động từ, động từ - đối tượng, động từ bị động - tác nhân, từ - giới từ đi kèm, danh từ - tính từ bổ nghĩa…

Các phụ thuộc theo loại này xác định các từ quan trọng trong câu và mối quan hệ giữa chúng. Trên cơ sở đó, luận án xây dựng nên các ràng buộc quan hệ bộ ba (constraint by triple patterns) trong câu truy vấn SPARQL.

Luận án cài đặt tác vụ phân tích gắn nhãn từ loại, cây cấu trúc cụm từ và phụ thuộc theo loại trong đó có tái sử dụng thư viện Stanford Parser.

3.4.3 Biểu diễn ngữ nghĩa cho câu hỏi

3.4.3.1 Mô hình biểu diễn ngữ nghĩa cho câu hỏi

Luận án đề xuất một mô hình biểu diễn ngữ nghĩa bao phủ hai dạng câu hỏi cơ bản: dạng câu hỏi có từ hỏi và dạng câu hỏi nghi vấn (yes/no). Mô hình biểu diễn ngữ nghĩa câu hỏi được trình bày chi tiết trong bảng 3.1 dưới đây.

Bảng 3.1. Mô hình biểu diễn ngữ nghĩa câu hỏi

Danh sách các biến: + Biến truy vấn (truy vấn số lượng, truy vấn đối tượng). + Biến thông thường.
Các ràng buộc cho các biến: + Ràng buộc nhãn của biến. + Ràng buộc quan hệ phụ thuộc giữa các biến. + Ràng buộc về số lượng
Ràng buộc cho các quan hệ phụ thuộc: + Ràng buộc AND/OR. + Ràng buộc thời gian.

Ý nghĩa của từng thành phần trong mô hình ngữ nghĩa trên như sau:

a) Danh sách các biến

Mỗi biến trong danh sách biến đại diện cho một từ (token) trong câu hỏi. Tên biến được đặt theo quy tắc: "chuỗi kí tự" + ID (ví dụ: ?x1, ?x2, …). Nhãn của biến chính là từ mà nó đại diện. Các biến được chia thành hai loại:

- Biến truy vấn: là những biến ẩn chứa thông tin cần trả về của câu truy vấn.
- Biến thường: là những biến còn lại.

Đối với dạng câu hỏi có từ hỏi, yêu cầu tồn tại ít nhất một biến truy vấn trong danh sách các biến, còn đối với dạng câu hỏi nghi vấn (yes/no) thì không tồn tại biến truy vấn. Tên của biến truy vấn được thêm dấu "?" phía trước để phân biệt với biến thường. Tác giả biểu diễn biến truy vấn dưới hai dạng: biến truy vấn số lượng đối với câu hỏi có từ hỏi là "how many" (biểu diễn trong danh sách biến là COUNT(?tên_biến)) và biến truy vấn đối tượng đối với những câu hỏi có từ hỏi là "who/what/which/where" (biểu diễn trong danh sách biến là ?tên_biến).

b) Các ràng buộc cho các biến

- Ràng buộc nhãn của biến: mỗi biến sẽ có nhãn là từ mà nó đại diện.
- Ràng buộc quan hệ phụ thuộc giữa các biến: mối quan hệ giữa hai biến ?subject và ?object sẽ được thể hiện bằng biến ?predicate dưới dạng bộ ba (?subject, ?predicate, ?object).
- Ràng buộc về số lượng: tác giả xác định ràng buộc về số lượng của một biến nào đó với một giá trị cụ thể thông qua các quan hệ: "=" (equal), ">" (morethan), ">= (moreOREqual)", "<" (lessthan), "<=" (lessOREqual).

c) Ràng buộc cho các quan hệ phụ thuộc

- Ràng buộc AND/OR: thể hiện việc các quan hệ phụ thuộc xảy ra đồng thời hay không nhất thiết đồng thời.
- Ràng buộc thời gian: giới hạn các quan hệ phụ thuộc (ví dụ như: trận đấu, chuyển nhượng cầu thủ…) xảy ra trong một khoảng thời gian nào đó.

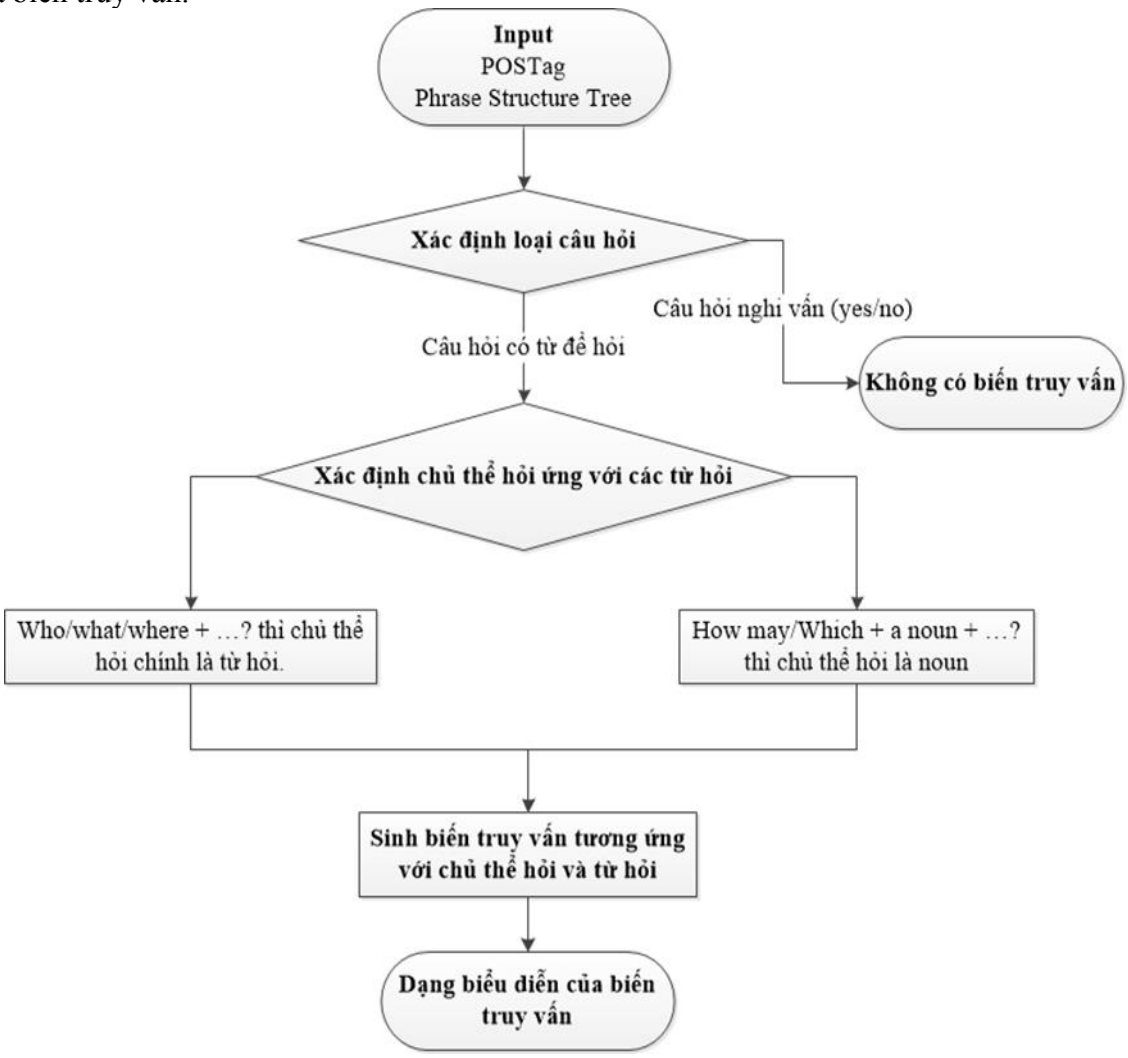
Để sinh ra được câu truy vấn SPARQL, luận án cần phải chuyển từ câu truy vấn dạng ngôn ngữ tự nhiên về mô hình biểu diễn ngữ nghĩa cho câu hỏi. Mô hình này là biểu diễn trung gian để sinh ra truy vấn SPARQL.

3.4.3.2 Chuyển từ cấu trúc ngữ pháp sang biểu diễn ngữ nghĩa

a) Xác định biến truy vấn

Hình 3.4 trình bày quy trình xác định biến truy vấn được luận án đề xuất. Như trình bày ở trên, sự tồn tại của biến truy vấn trong danh sách biến tùy thuộc vào dạng thức của câu truy vấn đầu vào. Nếu đầu vào là một câu hỏi nghi vấn thì không tồn tại biến truy vấn trong danh sách biến. Ngược lại, nếu đầu vào là một câu hỏi có từ hỏi, mô đun sẽ xác định chủ thể hỏi tương ứng với các từ hỏi. Đối với các từ hỏi là "who/what/where", mô đun sẽ xác định chủ thể hỏi chính là từ hỏi; còn đối với các từ hỏi là "how many/which" thì mô đun xác định chủ thể hỏi là danh từ đi sau từ hỏi, từ đó xác định được biến truy vấn.

Như đã trình bày ở tiểu mục mô hình biểu diễn ngữ nghĩa cho câu hỏi, tác giả chia biến truy vấn thành hai loại: biến truy vấn số lượng (cho từ hỏi "how many") và biến truy vấn đối tượng (cho từ hỏi "who/what/where/which"). Tùy thuộc vào loại từ hỏi, mô đun sẽ xác định được loại của biến truy vấn.

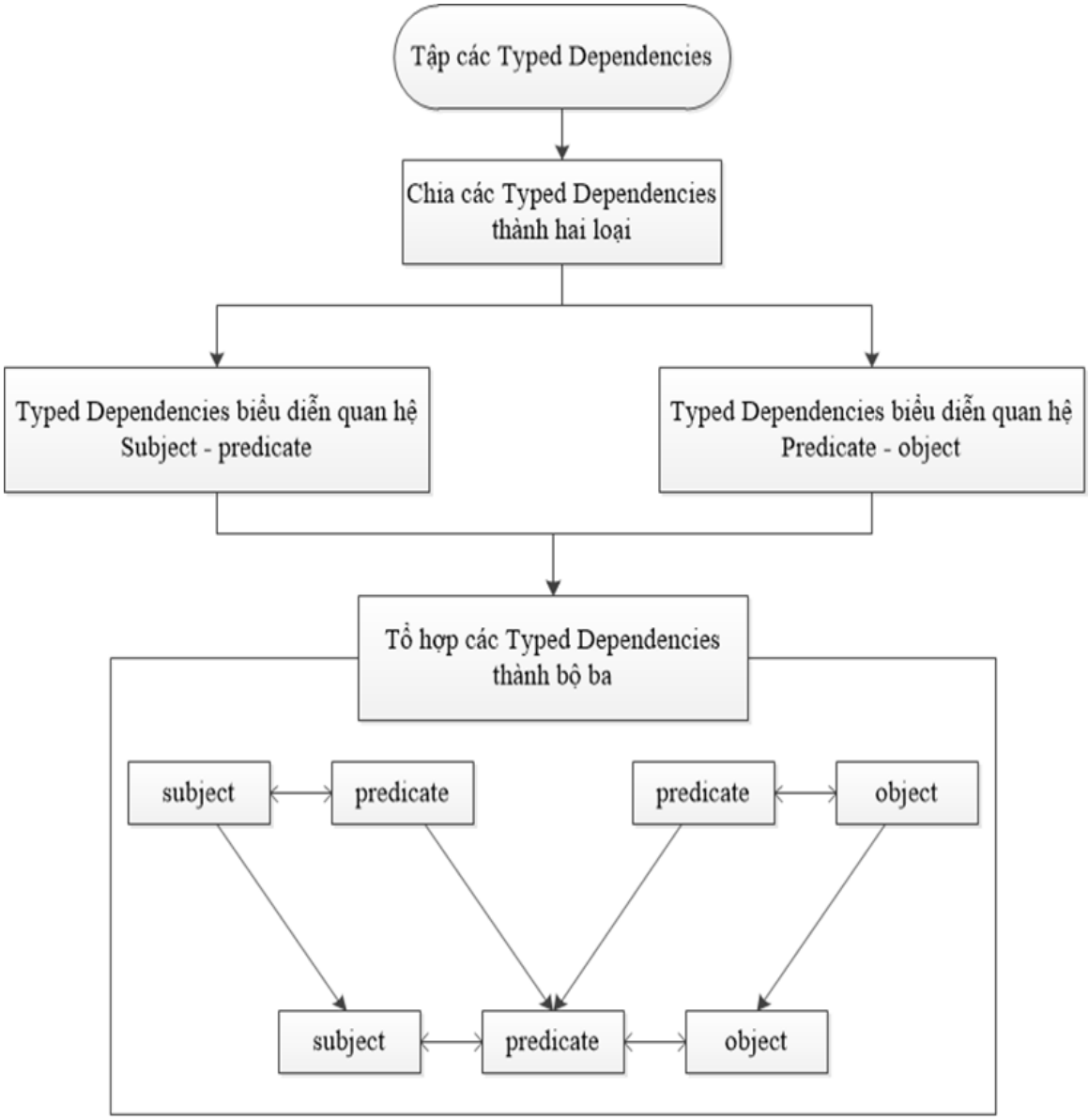


Hình 3.4 *Quy trình xác định biến truy vấn*

b) Xác định các biến thường và ràng buộc quan hệ phụ thuộc giữa các biến

Hình 3.5 minh họa quy trình xác định các biến thường và ràng buộc quan hệ phụ thuộc giữa các biến thực hiện trong luận án. Mỗi phụ thuộc theo loại là một bộ ba của: tên quan hệ, thành phần điều khiển và thành phần phụ thuộc. Từ các phụ thuộc theo loại thu được từ bước phân tích cú pháp, ta suy ra được các từ có quan hệ với nhau và mối quan hệ giữa chúng (dựa vào tên của phụ thuộc theo loại). Các từ này được đại diện bởi các biến, bao gồm cả biến truy vấn và biến thường.

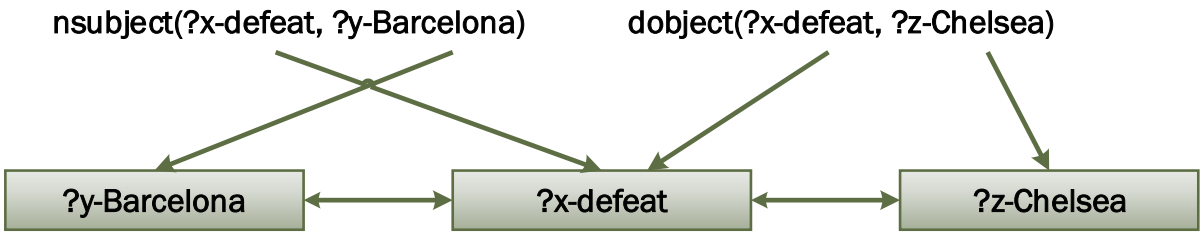
Như đã trình bày, một thành phần quan trọng của một câu truy vấn SPARQL là tập các bộ ba định nghĩa ràng buộc của câu hỏi. Tuy nhiên, mỗi quan hệ phụ thuộc giữa các biến xác định từ phụ thuộc theo loại mới chỉ là quan hệ bộ hai, các quan hệ phụ thuộc này thuộc một trong hai loại: subject - predicate và predicate – object. Vì thế, để sinh ra ràng buộc bộ ba cần kết hợp các phụ thuộc theo loại. Luận án thực hiện theo nguyên tắc sau: hai phụ thuộc theo loại biểu diễn hai loại quan hệ phụ thuộc khác nhau mà có chung vị ngữ (predicate) sẽ được xem xét để xây dựng nên quan hệ bộ ba có dạng (subject, predicate, object).



Hình 3.5 *Xác định các biến thường và ràng buộc quan hệ giữa các biến*

Ví dụ sau mô tả làm thế nào để thực hiện phương pháp cho dạng câu hỏi có ngữ nghĩa hỏi đơn giản S-V-O. Đối với câu chủ động "Did Barcelona defeat Chelsea?", mô đun phân tích cú pháp sẽ sinh ra được hai phụ thuộc theo loại chính là nsubject(?x-defeat, ?y-Barcelona) và

dobject(?x-defeat, ?z-Chelsea). Trong đó, nsubject biểu diễn quan hệ subject – predicate và dobject biểu diễn quan hệ predicate – object. Tổng hợp hai phụ thuộc theo loại này lại, ta thu được một quan hệ bộ ba (?y-Barcelona, ?x-defeat, ?z-Chelsea). Hình 3.6 dưới đây là hình ảnh trực quan mô tả việc kết hợp hai phụ thuộc theo loại thành quan hệ bộ ba.



Hình 3.6 *Phương pháp kết hợp hai phụ thuộc theo loại thành một quan hệ bộ ba*

Trong trường hợp câu bị động: "was Chelsea defeated by Barcelona?", mô đun phân tích cú pháp trả về hai phụ thuộc theo loại chính là nsubjectpass(?x-defeated, ?y-Chelsea) và agent(?x-defeated, ?z-Barcelona). Trong đó, nsubjectpass biểu diễn quan hệ predicate – object và agent biểu diễn quan hệ subject – predicate. Tương tự như trên, kết hợp hai phụ thuộc theo loại này lại ta được một quan hệ bộ ba (?z-Barcelona, ?x-defeated, ?y-Chelsea).

Đối với những dạng câu hỏi có ngữ nghĩa hỏi phức tạp như những câu có nhiều chủ thể, tân ngữ, mô đun phân tích cú pháp không chỉ sinh ra các phụ thuộc theo loại để biểu diễn quan hệ subject – predicate và predicate – object, mà còn sinh ra các phụ thuộc theo loại dạng conj_and(?x, ?y) hoặc conj_or(?x, ?y). Trong trường hợp này, hệ thống sinh ra hai bộ ba có cùng subject (hoặc cùng object) là ?x và ?y. Tùy thuộc vào loại liên kết "and" hay "or", câu truy vấn sẽ mang ý nghĩa "đồng thời xảy ra" hay "không nhất thiết đồng thời xảy ra" cả hai quan hệ bộ ba đó.

Ví dụ, đối với câu truy vấn "Which team defeated Chelsea and Barcelona?", mô đun phân tích cú pháp sẽ sinh ra các tập các phụ thuộc theo loại chứa nsubject(?x-defeated, ?y-team), dobject(?x-defeated, ?z-Chelsea) và conj_and(?z-Chelsea, ?t-Barcelona). Từ các phụ thuộc theo loại này, hệ thống sinh ra hai quan hệ đồng thời (?y-team, ?x-defeated, ?z-Chelsea) và (?y-team, ?x-defeated, ?t-Barcelona).

c) Xác định ràng buộc về số lượng

Để xác định các ràng buộc về số lượng, luận án vẫn dựa trên tập các phụ thuộc theo loại sinh ra từ bước phân tích cú pháp.

Luận án xem xét hai loại ràng buộc về số lượng: (1) ràng buộc so sánh số lượng của một đối tượng nào đó với một giá trị số cụ thể (ví dụ: Who scored more than 3 goals?) và (2) ràng buộc số lượng của một đối tượng nào đó là lớn nhất hay bé nhất (ví dụ: Who scored the most goals?, Which team conceded the least goals?).

Quy trình luận án đề xuất nhằm xác định ràng buộc về số lượng thuộc loại (1) được trình bày trong hình 3.7. Đối với loại (1), phụ thuộc theo loại num(?object, ?quantvalue) cho biết tồn tại một ràng buộc về số lượng cho đối tượng ?object dựa vào mối quan hệ với giá trị ?quantvalue. Để xác định quan hệ giữa đối tượng và giá trị số lượng này, luận án xem xét sự tồn tại của một phụ thuộc theo loại khác là quantmod(?quantvalue, "than"). Nếu phụ thuộc theo loại này không tồn tại, nghĩa là số lượng ?object bằng giá trị ?quantvalue. Ngược lại, nếu phụ thuộc theo loại này tồn tại, dựa vào giá trị của hai trường governor và dependent trong phụ thuộc theo loại mwe (?gov, ?dep) để xác định quan hệ bất đẳng thức giữa ?object và ?quantvalue (">", ">=", "<", "<=").

Đối với loại (2), luận án nhận thấy rằng có thể phát hiện cấu trúc "the most/least" + danh từ dựa trên hai phụ thuộc theo loại det(?object, "the") và amod(?object, ?dep). Nếu giá trị ?dep ở phụ thuộc theo loại admod là "most", nghĩa là đối tượng ?object có số lượng nhiều nhất, mô đun sinh ra ràng buộc themost(?object) trong mô hình ngữ nghĩa. Ngược lại, nếu giá trị này là "least", mô đun sinh ra ràng buộc theleast(?object).

Hình 3.7 *Quy trình xác định ràng buộc về số lượng loại (1)*

d) Xác định ràng buộc thời gian

Yếu tố về thời điểm luôn quan trọng đối với các sự kiện thể thao. Khác với các hệ thống QA khác, trong nghiên cứu này, luận án có tham vọng trả lời được các câu hỏi có ràng buộc về thời gian. Qua khảo sát, tác giả phân loại các câu hỏi loại này thành hai loại: ràng buộc gắn với một thời điểm và ràng buộc gắn với một khoảng thời gian.

Ví dụ:

- Loại 1: "today", "yesterday", "last sunday", "in 01/01/2015", …
- Loại 2: "last week", "last month", "this season", "this year", …

Một vấn đề đặt ra là phải xác định được thành phần của câu hỏi liên quan đến thời gian và chuyển đổi nó vào mô hình ngữ nghĩa. Như vậy công việc đầu tiên là định nghĩa cách biểu diễn ngữ nghĩa về mặt thời gian cho câu hỏi. Để làm điều này, trong mô hình ngữ nghĩa, tác giả định nghĩa một "Interval" gồm hai trường: Interval(BEGIN, END). Kiểu Interval thể hiện ràng buộc rằng thời điểm các sự kiện diễn ra phải nằm trong khoảng BEGIN và END.

Trong phạm vi nghiên cứu, luận án chỉ quan tâm đến đơn vị thời gian nhỏ nhất là ngày (ngày, tuần, tháng, năm). Nếu trong câu hỏi đầu vào có đề cập đến ngữ cảnh thời gian, bộ phân tích cú pháp sẽ sinh ra phụ thuộc theo loại prep_in(object, time_label). Trong đó time_label sẽ là nhãn thời gian thể hiện ngữ cảnh thời gian của cả câu hỏi. Nếu time_label là một ngày nào đó, BEGIN và END sẽ nhận cùng một giá trị. Còn nếu time_label là một tuần, tháng, năm hay mùa giải nào đó, luận án dựa vào thời điểm mà người dùng truy vấn để tính ra giá trị BEGIN và END.

Dưới đây là một số ví dụ biểu diễn câu hỏi trong mô hình ngữ nghĩa:
Ví dụ 1: "Which team defeated Chelsea in 08/05/2015?"

?x y z
x = "team" y = "defeated" z = "Chelsea" Triple1: (x, y, z)
Interval (08/05/2015, 08/05/2015)

Ví dụ 2: "Was Chelsea defeated by both Real Madrid and Barcelona in 2015?"

x y z t
x = "Chelsea" y = "defeated" z = "Barcelona" t = "Real Madrid" Triple1: (z, y, x) Triple2: (t, y, x)
AND(Triple1, Triple2) Interval (01/01/2015, 31/12/2015)

Từ mô hình biểu diễn ngữ nghĩa, mô đun tiếp theo sẽ sinh ra câu truy vấn SPARQL trung gian.

3.4.4 Sinh câu truy vấn SPARQL trung gian

Từ mô hình ngữ nghĩa của câu hỏi, giai đoạn xử lý kế tiếp là sinh câu truy vấn SPARQL trung gian được mô tả trong hình 3.8. Câu truy vấn trung gian này chỉ có khung chứa các biến, gồm hai thành phần chính là mệnh đề hỏi và mệnh đề điều kiện. Ngoài ra, đối với những câu hỏi dạng đặc biệt (câu hỏi có ràng buộc về số lượng, câu hỏi có ràng buộc thời gian) còn có thêm các mệnh đề ràng buộc khác.

Hình 3.8 *Quy trình sinh truy vấn SPARQL trung gian*

Quy trình sinh truy vấn SPARQL trung gian được trình bày trong hình 3.8. Các bước con trong quy trình này được trình bày chi tiết trong các tiểu mục dưới đây.

3.4.4.1 Xác định mệnh đề hỏi

Mệnh đề hỏi của câu truy vấn SPARQL gồm hai loại: mệnh đề SELECT hoặc mệnh đề ASK (tương ứng với hai dạng câu hỏi cơ bản: dạng câu hỏi có từ hỏi và dạng câu hỏi nghi vấn (yes/no)). Câu truy vấn dạng ASK chỉ trả về giá trị yes/no, được xác định khi trong mô hình ngữ nghĩa không có biến truy vấn. Câu truy vấn dạng SELECT sẽ trả về giá trị cụ thể cho biến truy vấn. Như đã phân biệt trong mô hình ngữ nghĩa, có hai loại biến truy vấn: biến truy vấn số lượng (COUNT(?x)) và biến truy vấn đối tượng (?x). Nếu trong danh sách biến chỉ chứa biến truy vấn số lượng (và biến thường) thì mệnh đề hỏi sẽ là "SELECT COUNT(?x)". Còn nếu trong danh sách biến chỉ chứa biến truy vấn đối tượng (và biến thường) thì mệnh đề hỏi sẽ là "SELECT ?x".

3.4.4.2 Xây dựng mệnh đề điều kiện – Mệnh đề WHERE

a) Sinh các bộ ba và biểu diễn mối quan hệ của chúng dựa vào ràng buộc AND/OR

Mệnh đề WHERE chứa các mẫu bộ ba là các bộ ba RDF ở dạng {<?subject> <?predicate> <?object>}. Các bộ ba này được xây dựng dựa trên các quan hệ bộ ba (?subject, ?predicate, ?object) trong mô hình ngữ nghĩa. Một quan hệ bộ ba trong mô hình ngữ nghĩa sẽ sinh ra một bộ ba trong mệnh đề WHERE. Ví dụ, nếu trong mô hình ngữ nghĩa có chứa bộ ba (?x, ?y, ?z), mô đun sẽ sinh ra một bộ ba có dạng: {<?x> <?y> <?z>}. Sự kết hợp các bộ ba này trong mệnh đề WHERE (kết hợp đồng thời hoặc không nhất thiết đồng thời) tùy thuộc vào ràng buộc AND/OR trong mô hình ngữ nghĩa. Nếu trong mô hình ngữ nghĩa có hai quan hệ bộ ba: bộ ba 1: (?x1, ?y1, ?z1) và bộ ba 2: (?x2, ?y2, ?z2) và tồn tại ràng buộc AND(Bộ_Ba_1, Bộ_Ba_2) thì hai bộ ba này sẽ được chuyển thành 2 bộ ba RDF và đi cùng nhau trong mệnh đề WHERE như sau:

```
 {
        <?x1> <?y1> <?z1>.
        <?x2> <?y2> <?z2>.
 }.
```

Mặt khác, nếu tồn tại ràng buộc OR(Bộ_Ba_1, Bộ_Ba_2), tác giả sẽ biểu diễn hai quan hệ này dưới dạng hợp như sau:

```
{
{<?x1> <?y1> <?z1>} UNION {<?x2> <?y2> <?z2>}.
}.
```

Mặc định, nếu không tồn tại ràng buộc AND/OR thì các bộ ba sẽ tuân theo ràng buộc AND.

b) Sinh các ràng buộc về số lượng

Luận án biểu diễn các ràng buộc về số lượng trong mô hình ngữ nghĩa bằng mệnh đề GROUP BY với các mệnh đề phụ bổ sung: HAVING và ORDER.

Mệnh đề HAVING

Mệnh đề HAVING dùng để biểu diễn các ràng buộc so sánh số lượng của một đối tượng nào đó với một giá trị số cụ thể. Giả sử trong mô hình ngữ nghĩa có ràng buộc moreORequal(?object, 3), tác giả sẽ biểu diễn trong câu truy vấn SPARQL như sau:

```
GROUP BY ?object
HAVING ( COUNT(?object) >= 3 ).
```

Mệnh đề ORDER:

Mệnh đề ORDER dùng để biểu diễn các ràng buộc số lượng của một đối tượng nào đó là lớn nhất hay bé nhất. Giả sử trong mô hình ngữ nghĩa có ràng buộc themost(?object), tác giả sẽ biểu diễn trong câu truy vấn SPARQL như sau:

```
GROUP BY ?object
ORDER BY DESC(COUNT(?object)) OFFSET 0 LIMIT 1.
```

Từ khóa "DESC" thể hiện rằng các kết quả truy vấn sẽ được sắp xếp giảm dần theo giá trị COUNT(?object) (ngược lại với "ASC" là sắp xếp tăng dần, dùng trong ràng buộc theleast(?object)). Ràng buộc "OFFSET 0 LIMIT 1" trong ví dụ này nghĩa là chỉ đưa kết quả ?object có giá trị COUNT(?object) lớn nhất.

c) Biểu diễn ràng buộc thời gian trong câu truy vấn SPARQL

Khác với các ràng buộc đã xét ở trên (ràng buộc giá trị nhãn cho biến, ràng buộc quan hệ phụ thuộc giữa các biến, ràng buộc về số lượng, chúng đều là những ràng buộc cho các biến hoặc là ?subject, hoặc là ?object), ràng buộc thời gian được xem xét trong nghiên cứu này là ràng buộc cho các quan hệ phụ thuộc giữa các biến, nghĩa là ràng buộc thời gian cho các quan hệ bộ ba. Để làm được điều này, luận án sử dụng mô hình NamedGraph để có thể gom nhóm các quan hệ bộ ba lại trong một đồ thị ?graph như sau:

```
?graph
{
        // các RDF triple
}
```

Sau đó, luận án định nghĩa ràng buộc thời gian cho các bộ ba RDF thông qua việc định nghĩa ràng buộc thời gian cho đồ thị ?graph như sau:

```
?g <http://bk.sport.owl#hasTime> ?t.
?t rdf:type time:Instant.
?t time:inXSDDateTime ?instantDate.
FILTER (?instantDate >= "BEGIN"^^<xsd:dateTime> && ?instantDate <= "END"^^<xsd:dateTime>).
```

Trong đó **BEGIN** và **END** là hai giá trị được xác định từ ràng buộc Interval (BEGIN, END) trong mô hình ngữ nghĩa của câu hỏi.

Ví dụ 1: với câu hỏi đầu vào "Which team defeated Chelsea in 08/05/2015?", luận án biểu diễn nó trong mô hình ngữ nghĩa như ví dụ 1 của tiểu mục 4.4.3.2 d). Dựa vào mô hình ngữ nghĩa này, hệ thống sinh ra câu truy vấn SPARQL trung gian như sau:

```
PREFIX rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
PREFIX time: <http://www.w3.org/2006/time#>
SELECT ?x
WHERE
{
    ?graph
    {
        ?x ?y ?z.
}
?g <http://bk.sport.owl#hasTime> ?t.
?t rdf:type time:Instant.
?t time:inXSDDateTime ?instantDate.
FILTER (?instantDate >= "2015-05-08"^^<xsd:dateTime> && ?instantDate <= "2015-05-08"^^<xsd:dateTime>).
```

}

Để câu truy vấn SPARQL trung gian trên trở thành câu truy vấn SPARQL hoàn chỉnh, hệ thống cần phải xác định giá trị và kiểu cho các biến ?x, ?y, ?z. Công việc này sẽ được luận án trình bày trong tiểu mục tiếp theo: mô đun xác định thực thể, khái niệm và vị từ.

3.4.5 Xác định thực thể, khái niệm và vị từ

Dựa vào các ràng buộc giá trị nhãn của biến trong mô hình biểu diễn ngữ nghĩa, mô đun "xác định thực thể, khái niệm và vị từ" tính toán kiểu và giá trị cho các biến này. Mô đun này sẽ ánh xạ từng nhãn của biến vào cơ sở tri thức và ontology để xác định xem các nhãn đó tương ứng với thực thể, lớp hay thuộc tính nào.

Sau khi tích hợp ontology BKSport vào ontology PROTON và làm giàu cơ sở tri thức của KIM, hệ thống KIM được mở rộng và tùy chỉnh đã sẵn sàng được sử dụng để nhận dạng những thực thể có tên trong lĩnh vực thể thao xuất hiện trong câu hỏi đầu vào. Mỗi biến trong mô hình ngữ nghĩa của câu hỏi đều có một nhãn riêng tương ứng. KIM dựa vào các nhãn này để chú thích ngữ nghĩa cho các biến. Hầu hết các biến có nhãn là một danh từ riêng thường tương ứng với một thực thể trong cơ sở tri thức. Một biến nếu được nhận dạng là một thực thể trong cơ sở tri thức thì nó được thay thế bằng URI của thực thể. Trường hợp một nhãn là danh từ riêng mà không được nhận dạng bởi KIM, hệ thống sẽ bổ sung vào câu truy vấn SPARQL trung gian một câu lệnh lọc FILTER để ràng buộc giá trị nhãn cho biến. Chi tiết của tác vụ nhận dạng thực thể có tên đã được trình bày ở chương 2, do đó mục này tập trung trình bày 2 tác vụ là xác định khái niệm (lớp) và vị từ (thuộc tính).

3.4.5.1 Nhận dạng các lớp

Các biến đóng vai trò là subject hoặc object trong các ràng buộc bộ ba nhưng không phải là thực thể thì luận án sẽ xây dựng ràng buộc lớp cho biến đó. Luận án tiến hành nhận dạng lớp cho biến theo các bước như sau. Đầu tiên, luận án xây dựng một danh sách gồm hai trường: trường thứ nhất là URI của tất cả các lớp có trong ontology và trường thứ hai là nhãn tương ứng của lớp đó. Sau đó, luận án dùng Wordnet để tìm các từ đồng nghĩa với các nhãn của từng URI trên, tạo ra một tập các từ đại diện cho mỗi URI. Hệ thống tiến hành kiểm tra nhãn của mỗi biến thuộc về tập từ đại diện nào, từ đó hệ thống xác định được URI tương ứng với biến và bổ sung vào câu truy vấn SPARQL một mẫu bộ ba (triple pattern) có cú pháp <tên_biến> <rdf:type> <URI_của_class> nhằm xác định kiểu của biến.

3.4.5.2 Nhận dạng thuộc tính

Việc nhận dạng thuộc tính của các biến đóng vai trò vị ngữ trong các bộ ba cũng tiến hành tương tự như việc nhận dạng các lớp. Tuy nhiên, có một số vấn đề khác nảy sinh cần phải xử lý riêng đối với quá trình này. Luận án nêu ra một số vấn đề và cách giải quyết:

a) Xử lý vấn đề một nhãn của biến vị ngữ tương ứng với nhiều thuộc tính trong ontology

Khi nhãn của một biến nào đó là một động từ và động từ này được nhận dạng thuộc nhiều tập từ đại diện của các URI khác nhau, khi đó nó sẽ sinh ra nhiều vị ngữ thỏa mãn (tuy nhiên, thường thì chỉ có một vị ngữ là đúng ý nghĩa của câu hỏi). Nguyên nhân là do nhãn của các thuộc tính trong ontology có thể được hợp thành bởi động từ và giới từ đi kèm với nó. Tuy nhiên, một động từ lại có thể đi kèm với nhiều giới từ khác nhau (ví dụ: play with, play for). Do đó, một nhãn của biến là một động từ có thể được nhận dạng thuộc nhiều lớp thuộc tính (do thiếu thông tin về giới từ). Cách giải quyết là nếu một động từ được nhận dạng thuộc nhiều tập từ đại diện khác nhau, dựa vào phụ thuộc theo loại **prep_"preposition"(?verb, ?object)** hệ thống lấy ra được giới từ của động từ đó. Nhãn của biến ?predicate bây giờ sẽ bao gồm động từ và giới từ đi kèm, giúp ta xác định duy nhất một kết quả.

b) Xử lý vấn đề vị ngữ là động từ "to be"

Khi vị ngữ là động từ "to be", thông tin cần truy vấn ở đây không chỉ là URI của biến truy vấn mà còn là định nghĩa (mô tả) của URI đó. Trong ontology BKSport, để mô tả một URI, tác

giả sử dụng thuộc tính bksport:hasAbstract. Do đó, khi một biến đóng vai trò vị ngữ mà có nhãn là động từ "to be", hệ thống sẽ thay thế nó bằng thuộc tính bksport:hasAbstract.

c) Xử lý vấn đề với vị ngữ là động từ "happen"

Khi người dùng muốn biết những sự việc diễn ra xoay quanh một (hay một vài) đối tượng, họ thường sử dụng động từ "happen". Trong ontology BKSport, tác giả tự định nghĩa thuộc tính **bksport:happen**. Đây là một thuộc tính ở mức cao, tổng quát, diễn tả một điều gì đó xảy ra. Các thuộc tính hành động ở mức thấp, cụ thể hơn sẽ là thuộc tính con của nó. Thông tin mà người dùng cần chính là những thuộc tính con này. Do vậy, luận án tạo một biến ?predicate khác thay thế biến cho thuộc tính **bksport:happen**, đồng thời, biến ?predicate này phải là thuộc tính con của thuộc tính **bksport:happen**. Điều kiện này được thỏa mãn bằng cách thêm vào câu truy vấn SPARQL một bộ ba dạng **"?predicate rdfs:subPropertyOf bksport:happen"**.

Ví dụ 1: với câu hỏi đầu vào "Which team defeated Chelsea in 08/05/2015?", mô hình ngữ nghĩa sinh ra được biểu diễn như ví dụ 1 của tiểu mục 3.4.3.2 d). Các ràng buộc về nhãn của các biến như sau:

x = "team"

y = "defeated"

z = "Chelsea"

Mô đun xác định thực thể, khái niệm và vị từ sẽ xác định được giá trị và kiểu của các biến x, y, z như sau:

Type(x) = http://bk.sport.owl#team

URI(y) = http://bk.sport.owl#defeat

URI(y) = http://bk.sport.owl#Chelsea

3.4.6 Sinh truy vấn SPARQL hoàn chỉnh

Sau mô đun xác định thực thể, khái niệm và vị ngữ, tất cả các biến trong mô hình ngữ nghĩa đã được xác định. Công việc sinh truy vấn SPARQL hoàn chỉnh đơn giản chỉ là thay thế các biến trong câu truy vấn SPARQL trung gian bằng các URI tương ứng.

Ví dụ 1: Kết hợp câu truy vấn trung gian trong ví dụ 1 của tiểu mục 3.4.4.2 c) và giá trị và kiểu của các biến trong ví dụ 1 của tiểu mục 3.4.5.3 c), hệ thống sinh ra câu truy vấn SPARQL hoàn chỉnh như sau:

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX bksport: <http://bk.sport.owl#>
SELECT ?x
WHERE
{
    ?graph
    {
        ?x <http://bk.sport.owl#defeat> <http://bk.sport.owl#Chelsea>.
}
```

```
?x rdf:type <http://bk.sport.owl#team>

?g <http://bk.sport.owl#hasTime> ?t.

?t rdf:type time:Instant.

?t time:inXSDDateTime ?instantDate.

FILTER (?instantDate >= "2015-05-08"^^<xsd:dateTime> && ?instantDate <= "2015-05-08"^^<xsd:dateTime>).

}
```

3.5 Thử nghiệm và đánh giá

3.5.1 Kịch bản thử nghiệm và kết quả

Trong nghiên cứu này, luận án chỉ tập trung trình bày hệ thống con chuyển đổi từ câu truy vấn của người dùng sang câu truy vấn SPARQL chứ chưa đề cập đến việc sử dụng truy vấn SPARQL để trả về câu trả lời từ kho tri thức ngữ nghĩa. Vì thế, trong phần này, luận án chỉ tiến hành đánh giá sự chính xác của việc chuyển đổi câu truy vấn. Tất cả các thực nghiệm được thực hiện trên máy tính Intel Core i7, CPU 2.30 GHz với RAM 8GB, hệ điều hành Microsoft Windows Server 2008. Các thuật toán được cài đặt bằng ngôn ngữ lập trình Java, sử dụng thư viện xử lý ngôn ngữ tự nhiên Stanford NLP version 1.3.5.

Để đánh giá được hệ thống, luận án tiến hành đánh giá trên từng câu truy vấn được chuyển đổi từ bộ dữ liệu câu hỏi tự nhiên đầu vào. Việc xác định một câu truy vấn SPARQL được sinh ra tự động bởi hệ thống là đúng hay sai có thể thực hiện bởi một người có kiến thức về truy vấn SPARQL. Để làm điều này, người đánh giá sẽ xem liệu câu truy vấn SPARQL sinh ra có thể hiện được đầy đủ và chính xác thông tin mà được thể hiện trong câu truy vấn ngôn ngữ tự nhiên đầu vào hay không. Tuy nhiên, để có thể đánh giá "mức độ đúng" của một câu truy vấn SPARQL là một điều khó. Điều này là do một câu truy vấn SPARQL được cấu thành bởi nhiều thành phần, mỗi thành phần của nó lại đóng vai trò khác nhau. Một câu truy vấn SPARQL gồm có 3 loại mệnh đề chính, mỗi loại mệnh đề được cấu thành bởi các thành phần con và thành phần đơn vị chính là các biến:

1. Mệnh đề hỏi: có các thành phần con là các biến hỏi (cũng chính là các thành phần đơn vị).
2. Mệnh đề WHERE: có các thành phần con là các bộ ba, mỗi bộ ba được cấu thành từ 3 thành phần đơn vị (subject, predicate, object) là các biến (các biến này đã được nhận dạng hoặc chưa).
3. Mệnh đề ràng buộc khác (mệnh đề ràng buộc thời gian, mệnh đề ràng buộc số lượng…): cấu thành từ các câu lệnh, mỗi câu lệnh đều chứa các thành phần đơn vị là các biến.

Ví dụ, phân tích một câu truy vấn có dạng như sau:

```
SELECT ?x1 ?x3

WHERE

{

  ?x1 ?x2 ?x3.

}

GROUP BY ?x1

ORDER BY DESC(COUNT(?x3)) OFFSET 0 LIMIT 1
```

Theo định nghĩa trên, ta thấy rằng:

- Mệnh đề hỏi: gồm hai thành phần đơn vị là ?x1 và ?x3

- Mệnh đề WHERE: có một bộ ba duy nhất, bộ ba này được cấu thành từ 3 thành phần đơn vị: ?x1, ?x2 và ?x3.
- Mệnh đề ràng buộc số lượng: bao gồm hai câu lệnh, câu lệnh thứ nhất chứa thành phần đơn vị ?x1, câu lệnh thứ hai chứa thành phần đơn vị ?x3.

Để đo lường được độ chính xác của một câu truy vấn, trước tiên ta phải đo được độ chính xác của từng loại mệnh đề. Để làm được điều này, luận án dựa vào các thành phần đơn vị. Tác giả định nghĩa một "thành phần đơn vị đúng" là một biến thỏa mãn một trong các điều kiện sau:

- Đã được nhận dạng (tương ứng với một URI)
- Đã được xác định rõ kiểu.
- Đã được xác định rõ ràng buộc về giá trị nhãn.

Luận án không đơn thuần đánh giá độ chính xác của một câu truy vấn sinh ra bởi hệ thống chỉ dựa trên số lượng các thành phần xác định đúng với ý muốn mà còn dựa trên độ quan trọng của mỗi thành phần. Để làm được điều này, luận án đánh trọng số cho từng loại mệnh đề trong câu truy vấn dựa vào quan điểm của tác giả về mức độ quan trọng của nó. Gọi w_i là trọng số của mệnh đề i trong câu truy vấn, luận án đánh trọng số như sau:

- $w_i = 3$, ứng với mệnh đề hỏi
- $w_i = 2$, ứng với mệnh đề WHERE
- $w_i = 1$, ứng với các mệnh đề ràng buộc khác.

Gọi n_i là số thành phần đơn vị đúng của loại mệnh đề i, N_i là số thành phần đơn vị cần xác định của loại mệnh đề i trong câu truy vấn được viết bởi chuyên gia, khi đó n_i/ N_i sẽ là độ chính xác của mệnh đề i. Một trường hợp đặc biệt đối với mệnh đề hỏi "SELECT *" mà không có biến hỏi cụ thể, luận án mặc định gán cho mệnh đề hỏi độ chính xác là 0.5.

Cuối cùng, luận án xác định công thức tổng quát để đo độ chính xác của một câu truy vấn q sinh ra bởi hệ thống như sau:

$$Precision(q) = b \times \frac{\sum_{i=1}^{M} \left(a_i \times w_i \times \frac{n_i}{N_i}\right)}{\sum_{i=1}^{M} (a_i \times w_i)} \quad (3.1)$$

Trong đó:

- b nhận giá trị 0 hoặc 1:
  - b = 0 nếu xác định sai loại mệnh đề hỏi (SELECT hay ASK) hoặc xác định sai tất cả các biến hỏi. Các biến hỏi quyết định đối tượng cần truy vấn là gì, vì vậy nếu tất cả các biến hỏi xác định sai thì precision = 0.
  - b = 1 trong các trường hợp còn lại.
- M là số loại mệnh đề có trong câu truy vấn viết bởi chuyên gia.
- $a_i = 1$ nếu mệnh đề loại i tồn tại trong câu truy vấn sinh bởi hệ thống, $a_i = 0$ nếu ngược lại.

Dựa vào những khảo sát về nhu cầu thông tin trong lĩnh vực thể thao của người đọc, luận án xây dựng một tập các câu hỏi nhằm đánh giá hệ thống mà luận án đã phát triển. Một phần của tập câu hỏi này được trình bày trong bảng 3.2 dưới đây. Mỗi câu hỏi trong tập câu hỏi này đại diện cho một mẫu câu hỏi khác nhau như luận án đã phân loại ở tiểu mục 3.3.1 Phân loại câu hỏi. Theo đó, tập dữ liệu thử nghiệm của luận án gồm 41 câu hỏi. Luận án đưa các câu hỏi này qua hệ thống đã xây dựng để tự động chuyển đổi về dạng truy vấn SPARQL, sau đó sử dụng công thức đề xuất trên để đánh giá độ chính xác cho từng câu truy vấn sinh ra. Kết quả độ đo của từng câu được thể hiện ở cột "Precision" của bảng. Tính cho toàn bộ tập câu hỏi, luận án sử dụng công thức tính trung bình cho độ đo của tất cả câu hỏi thử nghiệm. Kết quả nghiên cứu thu được là 91.89%.

Bảng 3.2. *Một phần của tập các câu hỏi để đánh giá hệ thống đề xuất*

ID	Question	Precision
	***** Definition question**	
1	Who is Lionel Messi?	1
	***** Yes/no question**	
2	Was Chelsea defeated by Barcelona last year?	1
3	Did Barcelona defeat Chelsea?	1
4	Did Wayne Rooney dispute with Alex Ferguson yesterday?	1
	***** Predicative question**	
5	Which team defeated Chelsea?	1
6	Which team defeated Chelsea this season?	0.83
7	Which event relates to Lionel Messi?	1
8	Which team did Lionel Messi transfer to?	1
	***** Opinion question**	
12	What did Lionel Messi say about Chelsea?	1
	***** Phrase-verb**	
13	news about Chelsea	1
	***** Quantity question**	
14	How many clubs defeated Chelsea?	1
	***** Comparative, superlative question**	
15	Who won the most games this year?	1
16	Who won more than 1 games this year?	1
	***** Association question:**	
20	What is the result of the match between Chelsea and Barcelona?	1
21	What happened between Chelsea and Barcelona?	1
	***** Multi-subject, multi-object question**	
22	Which team defeated Chelsea and Barcelona?	1
23	Which team defeated Chelsea and Barcelona in 2014?	1
24	Was Barcelona defeated by Manchester United and Chelsea this year?	1

3.5.2 Nhận xét và đánh giá

Vì hệ thống của luận án bao gồm nhiều bước nhỏ, sự không hoàn hảo trên kết quả của mỗi bước góp phần làm giảm độ chính xác của việc chuyển đổi câu truy vấn. Trong nghiên cứu này, luận án trình bày và sau đó phân tích một số khó khăn chính đến từ mỗi bước và một số nguyên nhân gây ra lỗi. Hướng khắc phục các hạn chế này sẽ được trình bày trong nội dung về nghiên cứu trong tương lai.

3.5.2.1 Phân tích cú pháp

Phân tích cú pháp là một trong những mô đun đầu tiên trong quy trình xử lý của hệ thống và đóng một vai trò quan trọng, ảnh hưởng đến hoạt động của các mô đun sau này như: nhận dạng thực thể, xác định các ràng buộc quan hệ bộ ba, ràng buộc ngữ cảnh thời gian… Như vậy, nếu bộ phân tích cú pháp hoạt động sai thì sẽ ảnh hưởng nghiêm trọng đến độ chính xác của cả quá trình chuyển đổi. Bộ phân tích cú pháp được sử dụng trong hệ thống là Stanford Parser (version 1.3.5). Đối với bộ câu hỏi trong tập câu hỏi thử nghiệm của luận án, bộ phân tích hoạt động đúng đối với hầu hết các câu hỏi. Điều này có được một phần là nhờ giai đoạn tiền xử lý, ở đó một số câu hỏi đầu vào không chuẩn đã được chuyển đổi sang dạng thức tương đương phù hợp nên bộ phân tích cú pháp vẫn phân tích đúng. Tuy nhiên, hệ thống vẫn không tránh khỏi một số trường hợp bị phân tích sai như: “Did Manchester United and Chelsea defeat Barcelona in

2014?". Với câu hỏi đầu vào trên, bộ phân tích cú pháp sinh ra các quan hệ phụ thuộc sai như: nn(defeat-6, Chelsea-5) và conj_and(United-3, defeat-6). Trong khi nếu phân tích đúng, bộ phân tích cú pháp cần trả ra kết quả rằng: nsubj(defeat-6, Chelsea-5) và conj_and(United-3, Chelsea-5). Đây là một hạn chế của phương pháp xử lý ngôn ngữ tự nhiên. Tuy nhiên, điều này có thể cải thiện bởi các phiên bản xử lý ngôn ngữ tự nhiên mới tốt hơn.

3.5.2.2 Nhận dạng quan hệ phụ thuộc bộ ba

Việc nhận dạng quan hệ phụ thuộc bộ ba chính là việc xác định 3 thành phần: subject, predicate, object dựa vào các phụ thuộc theo loại sinh ra ở bước phân tích cú pháp. Tuy nhiên, ngay cả trong trường hợp phân tích cú pháp đúng thì việc nhận dạng quan hệ bộ ba cũng có thể sai. Nguyên nhân dẫn đến điều này là do có những phụ thuộc theo loại sinh ra trong quá trình phân tích cú pháp câu hỏi đầu vào không nằm trong những phụ thuộc theo loại mà luận án xem xét, hoặc có thể do các phụ thuộc theo loại sinh ra biểu diễn các mối quan hệ phức tạp dẫn đến quá trình ghép các quan hệ bộ 2 thành quan hệ bộ 3 bị nhầm lẫn.

3.5.2.3 Nhận dạng khái niệm và vị từ

Bước nhận dạng khái niệm và vị từ là bước ánh xạ các từ/cụm từ trong câu hỏi tự nhiên vào các lớp và thuộc tính tương ứng trong cơ sở tri thức và ontology nhằm sinh ra ràng buộc về kiểu và giá trị cho các biến ở trong câu truy vấn SPARQL (không xét bước nhận dạng thực thể có tên, vì bước này hầu như không có trường hợp sai).

a) Nhận dạng khái niệm

Việc xác định sai kiểu (hay lớp) của biến thường hiếm khi xảy ra, tuy nhiên có xảy ra trường hợp không xác định được. Nguyên nhân là do ontology không chứa lớp tương ứng. Để giải quyết vấn đề này, cần phải xây dựng ontology phủ được hầu hết các khái niệm trong lĩnh vực đang xét.

b) Nhận dạng vị từ

Việc xác định sai giá trị của biến vị từ (biến thuộc tính là biến ?predicate trong quan hệ bộ ba <?subject> <?predicate> <?object>) cũng hiếm khi xảy ra. Trường hợp không xác định được giá trị của biến vị từ xảy ra là do có những thuộc tính tương ứng chưa được định nghĩa trong ontology. Ví dụ trường hợp câu hỏi đầu vào là "Which football player noun as CR7?" thì quan hệ "noun as" không nhận dạng được vì chưa định nghĩa quan hệ này trong ontology.

3.5.2.4 Xử lý nhãn thời gian

Tồn tại những nhãn thời gian mà hệ thống không thể xác định được ràng buộc như "this season". Vì có rất nhiều giải đấu, và mỗi mùa giải thì thời điểm bắt đầu và kết thúc lại không giống nhau, nên không thể xác định được ràng buộc.

3.5.2.5 Một số trường hợp đặc biệt chưa xử lý được

a) Trường hợp "Who is the best player in Chelsea?"

Mặc dù bộ phân tích cú pháp phân tích chính xác, tuy nhiên ontology lại không định nghĩa quan hệ thể hiện "cầu thủ chơi tốt nhất", vì thế không xác định được giá trị của biến vị từ.

b) Trường hợp: "Which player will leave Chelsea?"

Hệ thống còn chưa quan tâm đến "thì" của câu truy vấn đầu vào mà mặc nhiên coi như độc giả hỏi về những sự kiện đã xảy ra. Trên thực tế trong nghiên cứu sinh chú thích ngữ nghĩa, một số predicate biểu thị độ chắc chắn của kết quả đã được nghiên cứu trong chương 2. Do đó, trong tương lai gần lỗi này có thể được khắc phục nhờ biện pháp kết hợp với việc xử lý ràng buộc về thời gian.

3.6 Kết luận chương

Chương này đã trình bày hệ thống chuyển đổi câu truy vấn dạng ngôn ngữ tự nhiên về câu truy vấn SPARQL. Câu truy vấn ở dạng ngôn ngữ tự nhiên là đầu vào của hệ thống được xử lý

tự động qua nhiều mô đun con để sinh ra câu truy vấn SPARQL hoàn chỉnh. Câu truy vấn trước tiên sẽ được tiền xử lý, sau đó bộ phân tích cú pháp sẽ phân tích nó để nhận biết các thành phần ngữ pháp và mối quan hệ giữa các thành phần ngữ pháp đó, từ đó biểu diễn câu truy vấn dưới dạng mô hình ngữ nghĩa. Từ mô hình ngữ nghĩa, mô đun sinh truy vấn SPARQL trung gian sẽ tạo ra một khung truy vấn SPARQL chỉ chứa các biến. Cuối cùng, mô đun xác định thực thể, khái niệm và vị từ sẽ chú thích và sinh ràng buộc cho các biến trong khung truy vấn SPARQL trung gian bằng các URI trong ontology và cơ sở tri thức, sinh ra câu truy vấn SPARQL hoàn chỉnh.

Kết quả trên đã được tác giả công bố trong bài báo "Sport News Semantic Search with Natural Language Questions" được báo cáo tại hội nghị quốc tế *European Alliance for Innovation (EAI) International Conference on Industrial Networks and Intelligent Systems* (INISCOM 2018).

Dựa trên việc tiền xử lý và phân tích sâu cấu trúc ngữ pháp của câu truy vấn, nên hệ thống mà luận án đề xuất có khả năng xử lý được một số dạng câu truy vấn phức tạp như câu hỏi so sánh hơn, so sánh hơn nhất, câu hỏi có nhiều chủ ngữ, tân ngữ, câu hỏi có cấu trúc ngữ pháp không chuẩn tắc… Một số trường hợp do kết quả phân tích cú pháp sinh ra phức tạp, hay ngữ nghĩa của câu hỏi phức tạp mà hệ thống xử lý chưa đúng. Tuy nhiên, đối với những câu hỏi phức tạp như vậy, hệ thống cũng đã xử lý đúng phần nào. Qua thử nghiệm và đánh giá trên bộ câu hỏi gồm nhiều loại câu hỏi khác nhau cho thấy hệ thống đã đề xuất đạt độ chính xác cao.

Trong tương lai, những nghiên cứu tiếp theo sẽ tập trung cải thiện những trường hợp mà hệ thống hiện tại chưa xử lý đúng và hoàn thiện ontology BKSport để bao phủ được đầy đủ các khái niệm và các quan hệ có trong miền lĩnh vực thể thao, quan tâm đến "thì" của quan hệ là động từ nhằm nắm bắt ngữ nghĩa của câu truy vấn một cách chính xác hơn. Hệ thống này sẽ được tích hợp vào cổng thông tin thể thao BKSport đã được xây dựng và đang hoàn thiện, hỗ trợ cho việc tìm kiếm tin tức hiệu quả.

CHƯƠNG 4. GỢI Ý TIN TỨC DỰA TRÊN NGỮ NGHĨA CHO HỆ THỐNG TỔNG HỢP TIN TỨC THỂ THAO

Ngày nay, tin tức trên Internet đóng một vai trò quan trọng trong việc giúp mọi người tiếp cận các thông tin diễn ra hàng ngày trên thế giới. Tuy nhiên, số lượng tin tức trên Internet liên tục tăng gây khó khăn cho độc giả khi muốn tiếp cận một tin tức mà mình quan tâm. Để giải quyết vấn đề này, các hệ thống gợi ý tin tức đã được xây dựng. Có nhiều phương pháp gợi ý tin tức đã được nghiên cứu, hầu hết các phương pháp đều dựa trên một độ đo tương đồng nào đó (độ tương đồng giữa hai tin với nhau hoặc giữa sở thích cá nhân của độc giả và tin). Trong nghiên cứu này, luận án đề xuất và cài đặt thực nghiệm một phương pháp gợi ý tin tức thể thao dựa trên kết hợp độ tương đồng về ngữ nghĩa với độ tương đồng về nội dung của hai tin tức. Kết quả thử nghiệm cho thấy rằng khi kết hợp cả hai độ đo sẽ cho kết quả gợi ý tốt hơn khi chỉ dùng một trong hai độ đo.

4.1 Giới thiệu

Khi con người bắt đầu đọc tin tức trực tuyến ngày càng nhiều hơn thì việc tìm thấy các tin tức thú vị và hợp với các yêu cầu của họ đã trở thành một thách thức. Trong chương 2 và chương 3, luận án hướng đến việc cải thiện tính năng tìm kiếm cho hệ thống tổng hợp tin tức thể thao, tuy nhiên trong thực tế không phải lúc nào ta cũng biết rõ về tin tức mà mình muốn xem.

Các hệ thống gợi ý nói chung được xây dựng để giúp chúng ta dễ dàng tìm ra thông tin phù hợp nhất trên Internet. Không giống như các hệ thống tư vấn bằng công cụ tìm kiếm, nó mang đến thông tin cho người dùng mà không có bất kỳ nỗ lực tìm kiếm thủ công nào. Điều này đạt được bằng cách sử dụng những tương đồng giữa người dùng và mục tin. Có nhiều phương pháp để xây dựng một hệ thống gợi ý và những phương pháp này có thể được áp dụng cho nhiều lĩnh vực cụ thể như mua sắm (ví dụ như Amazon), phim ảnh (ví dụ như Netflix), và nhạc (ví dụ như Pandora Radio).

Các hệ thống gợi ý tin tức có mục đích đưa ra gợi ý về các bài viết phù hợp nhất cho độc giả, mà có cân nhắc đến dự đoán theo những ưu tiên và sở thích cá nhân của họ.

Gợi ý tin tức trong lĩnh vực thể thao là một trong những nhiệm vụ thách thức nhất, vì lĩnh vực thể thao khác hẳn với các lĩnh vực khác như âm nhạc, mua sắm, phim ảnh. Một ví dụ điển hình là tính thời sự và tính phổ biến của các tin tức thể thao thay đổi quá nhanh theo thời gian. Vì vậy, nếu chỉ tập trung vào giải quyết vấn đề tính tươi mới trong lĩnh vực tin tức sẽ khó hơn. Ngoài ra, một số tin tức cũng có thể được liên kết với nhau để giúp độc giả theo dõi tiếp các tin tức có liên quan đến tin tức mà họ đã đọc hoặc có quan tâm.

Có nhiều tiếp cận khác nhau cho bài toán gợi ý nói chung và gợi ý tin tức nói riêng, trong đó nổi bật là phương pháp gợi ý dựa trên sự tương đồng giữa các tin tức với nhau, còn được gọi là phương pháp dựa trên nội dung. Một tiếp cận khác dựa trên sự tương đồng giữa tin tức và sở thích cá nhân của người đọc [136]. Lọc cộng tác là kỹ thuật phổ biến được ứng dụng trong trường này. Tuy nhiên, chỉ tìm hiểu sở thích người đọc có thể là giải pháp không đầy đủ cho việc gợi ý tin tức, bởi vì độc giả có thể muốn đọc một tin tức khi không thực sự quan tâm đến chủ đề nhưng lại nghĩ rằng nó quan trọng. Ví dụ, họ muốn đọc tin về cuộc bầu cử ngay cả khi không hề quan tâm đến lĩnh vực chính trị.

Chương 4 của luận án này trình bày nghiên cứu khai thác khía cạnh ngữ nghĩa nhằm cải thiện khả năng hoạt động của hệ thống tổng hợp tin tức, giúp nó không những có tính năng tìm kiếm mà còn có cả chức năng khuyến nghị (gợi ý). *Mục tiêu được xác định là nâng cao hiệu quả của phương pháp dựa trên nội dung với ý tưởng kết hợp độ tương đồng nội dung với độ tương đồng ngữ nghĩa.*

Chương 4 tập trung trình bày tiếp cận lai (hybrid), kết hợp gợi ý dựa trên nội dung và dựa trên ngữ nghĩa cho hệ thống. Phương pháp này kế thừa các kết quả thu được trong các nghiên cứu trước đây như ontology và cơ sở tri thức trong lĩnh vực thể thao, các phương pháp nhận dạng thực thể có tên và trích rút các quan hệ ngữ nghĩa giữa các thực thể trong tin tức.

Các mục còn lại của chương 4 được tổ chức như sau. Mục 4.2 mô tả các nghiên cứu trước đây liên quan đến đo độ tương đồng về ngữ nghĩa giữa hai tin tức. Mục 4.3 trình bày chi tiết hơn về phương pháp đề xuất. Mục 4.4 trình bày các thử nghiệm đã được thực hiện bằng cách sử dụng cài đặt của các chuyên gia về gợi ý đề xuất và đánh giá những kết quả thu được. Sau đó, ưu điểm và nhược điểm của phương pháp này cũng như các biện pháp khắc phục và hướng nghiên cứu trong tương lai được kết luận trong mục 4.5.

4.2 Nghiên cứu liên quan

Như đã đề cập, các phương pháp gợi ý dựa trên lọc cộng tác sử dụng các sở thích của độc giả khác mà có sự tương đồng với sở thích trong quá khứ của một độc giả xác định. Tin tức được gợi ý cho độc giả đó là tin tức đã được đọc bởi nhiều người có sở thích tương đồng với sở thích của độc giả hiện tại nhất [44, 45, 46]. Tuy nhiên, tiếp cận này đòi hỏi thông tin về lịch sử đọc tin của rất nhiều người dùng. Vì vậy, luận án không theo tiếp cận này.

Nhiều nghiên cứu cho thấy rằng các hệ thống khuyến nghị dựa trên nội dung thường cố gắng giới thiệu các tin tức có độ tương đồng cao nhất với tin tức người đọc quan tâm. Xây dựng mô hình để tính độ tương đồng giữa các tin tức đóng vai trò quan trọng trong tiếp cận này.

Trong tiếp cận thuần túy dựa trên nội dung (content-based), độ tương đồng tin tức được tính toán dựa trên các thống kê từ vựng xuất hiện trong nội dung tin tức và hầu hết các tin được gợi ý chỉ tập trung vào chủ đề mục tiêu mà tin tức đang hướng tới. Ngược lại, trong tiếp cận dựa trên ngữ nghĩa (semantic-based) [137], độ tương đồng tin tức thường dựa trên cơ sở tri thức có sẵn để khai thác mối quan hệ ngữ nghĩa giữa các yếu tố xuất hiện trong những tin này. Vì vậy, những tin tức được gợi ý sẽ có khả năng mở rộng các chủ đề hơn so với cách tiếp cận dựa trên nội dung.

Theo truyền thống, nhiều nhà nghiên cứu về chủ đề gợi ý hướng nội dung (content-based recommenders) [47, 48] sử dụng các phương pháp trích rút thuật ngữ như TF-IDF (Term Frequency-Inverse Document Frequency). [138] kết hợp TF-IDF với phép đo độ tương đồng cosin để so sánh độ tương đồng giữa hai tài liệu. TF-IDF được sử dụng để đo độ quan trọng của một từ trong một tài liệu dựa trên tần xuất xuất hiện của nó trong toàn bộ tập dữ liệu tài liệu (hoặc tập dữ liệu). Sau khi tính giá trị TF-IDF cho mỗi từ trong tài liệu, chỉ số này được kết hợp với phép đo Cosine hoặc phép đo Jacard để tính độ tương đồng giữa hai tài liệu.

Giá trị TF-IDF của từ xuất hiện trong tài liệu được tính theo công thức sau:

$$TF\text{-}IDF_{ij} = TF_{ij} \times IDF_i \tag{4.1}$$

Trong đó:

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{4.2}$$

và

$$IDF_{ij} = log \frac{|D|}{|\{d: t_i \in d\}|} \tag{4.3}$$

n_{ij} là số lần xuất hiện của từ i trong tài liệu j, và $|D|$ là tổng số tài liệu trong tập dữ liệu, $|\{d: t_i \in d\}|$ là số tài liệu d trong tập D mà chứa từ i (t_i).

Sau đó, tài liệu được biểu diễn bằng một véc tơ V_i có chiều N (với N là kích thước của từ điển), giá trị của mỗi phần tử của vectơ là giá trị TF-IDF của từ. Nếu từ này trong từ điển không thuộc về tin tức, giá trị của phần tử tương ứng trong vectơ là 0.

Trong cách tiếp cận hướng dựa trên ngữ nghĩa, các nghiên cứu trước đây đã khám phá mối liên quan giữa các thành phần giữa hai tin với nhau để tính độ tương đồng ngữ nghĩa. Trong nghiên cứu được thực hiện bởi [49], một phép đo dựa trên việc khai thác cấu trúc phân loại của một ontology sinh học được đề xuất để xác định độ tương đồng về ngữ nghĩa giữa các cặp từ. Phương pháp được đề xuất bởi [50] khai thác yếu tố tương đồng giữa các thành phần (từ hoặc thực thể có tên) trong tin tức, từ đó tính độ tương đồng giữa hai tin tức. Để đo độ tương đồng giữa hai thành phần, phương pháp đề xuất của họ dựa trên:

- Cây từ điển WordNet khi các thành phần là từ - ký hiệu là sim_{SS}
- Phép đo PMI khi các thành phần là các thực thể có tên - ký hiệu là sim_{Bing}. Phương pháp này liên quan đến tần số thống kê sự xuất hiện của các thành phần và sự xuất hiện đồng thời giữa chúng

Công thức cuối cùng kết hợp hai độ đo sim_{SS} và sim_{Bing} để tính độ tương đồng ngữ nghĩa giữa hai tin như sau (α là tham số hiệu chỉnh):

$$sim_{BingSS} = \alpha \times sim_{Bing} + (1-\alpha) \times sim_{SS} \quad (4.4)$$

Ngoài ra khi khai thác mối quan hệ giữa các thành phần trong hai tin tức với nhau, [139] đưa ra một vài phương pháp gợi ý tin sử dụng tiếp cận theo hướng dựa trên nội dung. Tương tự như ở [50], công trình của họ hướng đến một hệ thống gợi ý cá nhân hóa (PRS). Tuy nhiên hồ sơ người dùng của độc giả cũng vẫn được xây dựng dựa trên những tin tức mà độc giả đã đọc, rồi tính độ tương đồng giữa hồ sơ người dùng với tin tức cũng giống như cách tính độ tương đồng giữa hai tin. Các phương pháp được trình bày trong nghiên cứu này sử dụng ontology và cơ sở tri thức để khai thác mối quan hệ ngữ nghĩa giữa các khái niệm (các lớp trong ontology). Thử nghiệm cho thấy phương pháp *Ranked Semantic Recommendation 2* là hiệu quả nhất trong các phương pháp. Tuy nhiên, những hạn chế còn tồn tại sẽ được luận án trình bày trong các mục sau, đồng thời phương pháp khắc phục cũng được đề xuất.

4.3 Độ tương đồng giữa các tin

Có hai phương pháp tiếp cận chính để tính độ tương đồng giữa các tin dạng văn bản, đó là hướng dựa trên nội dung và hướng dựa trên ngữ nghĩa. Mỗi phương pháp tiếp cận đều có ưu điểm và nhược điểm riêng. Với kỳ vọng khắc phục những hạn chế của từng phương pháp và giúp cho việc gợi ý hiệu quả hơn, luận án đã kết hợp hai phương pháp tiếp cận này bằng cách kết hợp độ đo tương đồng về nội dung và độ đo tương đồng về ngữ nghĩa.

4.3.1 Độ tương đồng về ngữ nghĩa

Để tính toán độ tương đồng ngữ nghĩa, luận án tập trung khai thác các quan hệ ngữ nghĩa lẫn nhau giữa các thành phần trong các tin tức. Các mối quan hệ này được xác định dựa trên ontology và cơ sở tri thức sẵn có trong hệ thống tổng hợp tin tức thể thao BKSport. Các thành phần trong tin bao gồm: các thực thể, các loại thực thể và các chú thích ngữ nghĩa được trích rút và phân tích. Các tiểu mục tiếp theo sẽ trình bày cách khai thác các thành phần này để tính độ tương đồng ngữ nghĩa giữa các tin. Sau đây là các pha thực hiện.

4.3.1.1 Quan hệ ngữ nghĩa giữa các thực thể

Để tính sự tương đồng giữa các mục tin, luận án đề xuất khai thác mối quan hệ giữa các thực thể. Tác giả mở rộng phương pháp *Ranked Semantic Recommendation 2* mà Frasincar và cộng sự đề xuất [139]. Trong phương pháp này, Frasincar và các cộng sự cũng sử dụng ontology và cơ sở tri thức để khai thác mối quan hệ giữa các thực thể, tuy nhiên phương pháp vẫn còn một số hạn chế như:

- Chỉ xem xét mối quan hệ trực tiếp giữa các thực thể mà chưa xét đến các mối quan hệ gián tiếp.
- Chưa xét đến độ quan trọng của các thực thể khi chúng xuất hiện ở các vị trí khác nhau trong tin tức như trong tiêu đề hay trong mô tả …

Để khắc phục những hạn chế trên, trong tiểu mục 4.3.1.1.a) dưới đây, luận án trình bày phương pháp để tính trọng số quan hệ giữa các thực thể dựa vào ontology và cơ sở tri thức. Tiếp theo, tiểu mục 4.3.1.1.b) trình bày sự kết hợp phương pháp thống kê sự đồng xuất hiện của các thực thể trong cùng một tin để xác định trọng số quan hệ giữa các thực thể. Tiểu mục 4.3.1.1.c) trình bày phương pháp sử dụng các trọng số quan hệ giữa các thực thể để xác định độ tương đồng về ngữ nghĩa giữa các mục tin.

a) Trọng số quan hệ giữa các thực thể dựa vào ontology và cơ sở tri thức

Trong nghiên cứu [140], nhóm tác giả trình bày các phương pháp để tính việc xếp hạng các liên kết ngữ nghĩa dựa vào đường đi ngữ nghĩa giữa hai thực thể để xác định trọng số quan hệ giữa các thực thể này. Các tác giả định nghĩa liên kết ngữ nghĩa và đường đi ngữ nghĩa như sau:

Định nghĩa: nếu hai thực thể $e_1$và e_n có thể được kết nối với nhau bằng một hoặc nhiều dãy $e_1, P_1, e_2, P_2, e_3, P_3, \dots, e_{n-1}, P_{n-1}, e_n$, e_n trong một đồ thị RDF; Ở đây e_i ($1 \le i \le n$) là các thực thể và P_j ($1 \le j \le n$) là các quan hệ trong ontology, thì ta nói có tồn tại mối quan hệ ngữ nghĩa giữa $e_1$và e_n.

Và kết quả là dãy $e_1, P_1, e_2, P_2, e_3, P_3, \dots, e_{n-1}, P_{n-1}, e_n$, là một đường đi ngữ nghĩa.

Ví dụ, trong cơ sở tri thức, ta có:

- <Lionel-Messi> <playFor> <Barcelona-FC>.
- <Luis-Suarez> < playFor > <Barcelona-FC>.

Khi đó, giữa hai thực thể *Lionel Messi* và *Luis Suarez* sẽ tồn tại đường đi ngữ nghĩa như sau:
<Lionel-Messi> → <playFor> → <Barcelona-FC> ← <playFor> ← <Luis-Suarez>

Như vậy, tồn tại mối quan hệ ngữ nghĩa giữa hai thực thể *Lionel Messi* và *Luis Suarez*.

Dựa vào các tính chất của đường đi ngữ nghĩa, luận án xác định một giá trị *xếp hạng đường đi* (path rank) thể hiện trọng số quan hệ giữa hai thực thể ở hai điểm đầu của một đường đi. Bởi vì giữa hai thực thể có thể tồn tại nhiều đường đi ngữ nghĩa, luận án chọn giá trị *xếp hạng đường đi* cao nhất để đại diện cho trọng số quan hệ. Trong [140], các tác giả sử dụng bốn đặc trưng của một đường đi ngữ nghĩa để tính giá trị *xếp hạng đường đi*. Chúng là bốn trọng số sau:

- *Trọng số gộp* (Subsumption Weight): dựa trên cấu trúc của ontology để xác định *trọng số thành phần* (component weight) cho từng thành phần (quan hệ và thực thể) trong đường đi, từ đó tính trọng số cho toàn bộ đường đi.
- *Trọng số độ dài đường đi* (Path Length Weight): được tính dựa trên độ dài đường đi.
- *Trọng số ngữ cảnh* (Context Weight): dựa trên việc xác định mỗi thành phần của đường đi thuộc vùng nào trong ontology. Mỗi vùng trong ontology sẽ có một trọng số riêng tùy thuộc vào sở thích của người dùng.
- *Trọng số tín nhiệm* (Trust Weight): được tính dựa trên trọng số của các thuộc tính trong ontology.

Khi áp dụng vào bài toán đặc thù là gợi ý tin tức về lĩnh vực bóng đá, tác giả thấy rằng hai *trọng số độ dài đường đi* và *trọng số tín nhiệm* là hai trọng số lớn nhất và phù hợp nhất. Vì lý do này, luận án chỉ quan tâm đến hai trọng số này khi tính toán giá trị *xếp hạng đường đi* của một đường đi ngữ nghĩa.

Trọng số xếp hạng dựa vào độ dài đường đi ngữ nghĩa (Path Length Weight)

Độ dài của một đường đi ngữ nghĩa $e_1, P_1, e_2, P_2, e_3, P_3, \dots, e_{n-1}, P_{n-1}, e_n$ là số thực thể và các quan hệ trong đường đi (không tính hai thực thể e_1 và e_n ở hai đầu mút). Chúng ta có thể thấy rằng khi hai thực thể có quan hệ gián tiếp với nhau thông qua càng nhiều thực thể và quan hệ, thì hai thực thể này có độ tương đồng càng thấp. Do đó, giá trị *xếp hạng đường đi* của một đường đi ngữ nghĩa phải tỉ lệ nghịch với độ dài của đường đi đó.

Trọng số xếp hạng đường đi ngữ nghĩa dựa vào độ dài của nó (Path Length Weight) được định nghĩa trong [140] như sau:

$$W_{length} = \frac{1}{length_{path}} \tag{4.5}$$

Trong đó: $length_{path}$ là độ dài của đường đi ngữ nghĩa.

Ví dụ, ta có hai đường đi ngữ nghĩa:

- P_1: <Lionel-Messi> → <playFor> → <Barcelona-FC> → <competeIn> → <La-Liga> ← <competeIn> ← <Real-Madrid> ← <playFor> ← <Sergio-Ramos>
- P_2 :<Lionel-Messi> → <playFor> → <Barcelona-FC> ← <playFor> ← <Luis-Suarez>

P_1 có độ dài đường đi là 7, ta tính được:

$$W_{length}(P_1) = \frac{1}{length_{path}} = \frac{1}{7}$$

P_2 có độ dài đường đi là 3, ta tính được:

$$W_{length}(P_2) = \frac{1}{length_{path}} = \frac{1}{3}$$

Từ đó, ta có thể thấy độ tương đồng giữa *Lionel Messi* và *Luis Suarez* cao hơn giữa *Lionel Messi* và *Sergio Ramos*.

Trọng số xếp hạng dựa vào quan hệ đường đi ngữ nghĩa (Path Relation Weight)

Có rất nhiều quan hệ được định nghĩa trong ontology, mỗi quan hệ thể hiện một ý nghĩa khác nhau. Do đó khi đóng vai trò liên kết hai thực thể, mỗi quan hệ có một trọng số khác nhau thể hiện sự liên quan giữa các thực thể. Một số quan hệ thể hiện sự liên quan mật thiết, một số quan hệ khác lại thể hiện sự liên quan yếu hơn. Ví dụ, ta có các bộ ba trong cơ sở tri thức như sau:

- <Luis-Enrique> <managerOf> <Barcelona-FC>.
- <Luis-Suarez> <cầu thủ> <Barcelona-FC>.

Ở đây, tồn tại hai quan hệ là quan hệ <*managerOf*> và quan hệ <*playFor*>. Ta có thể thấy rằng quan hệ <*managerOf*> thể hiện mối liên quan mật thiết hơn mối quan hệ <*playFor*> vì mỗi đội chỉ có một huấn luyện viên duy nhất tại một thời điểm. Tuy nhiên, lại có thể có rất nhiều cầu thủ. Do đó, ta đánh trọng số cho mối quan hệ <*managerOf*> cao hơn <*playFor*>. Và vì lý do này, từ hai bộ ba trên, ta kết luận <*Barcelona-FC*> có độ tương đồng với <*Luis-Enrique*> cao hơn <*Luis-Suarez*>.

Trọng số của các quan hệ nằm trong khoảng (0,1). Công thức tính trọng số xếp hạng đường đi ngữ nghĩa dựa vào các quan hệ có trong đường đi (*Path Relation Weight*) trong [140] như sau:

$$W_{predicate} = \prod_{p \in path} w_p \tag{4.6}$$

Trọng số quan hệ giữa hai thực thể dựa vào Ontology và cơ sở tri thức

Kết hợp hai trọng số W_{length} và $W_{predicate}$ bằng một cặp hệ số α_{wl} và α_{wp}, ta tính được *path- rank* của đường đi ngữ nghĩa như sau:

$$W_{path} = \frac{W_{length} \times \alpha_{wl} + W_{predicate} \times \alpha_{wp}}{\alpha_{wl} + \alpha_{wp}} \tag{4.7}$$

Trong đó, giá trị của các hệ số α_{wl} và α_{wp} có tổng bằng 1.0 và được tinh chỉnh tùy theo quan điểm đánh giá về độ ảnh hưởng của hai trọng số. Giá trị W_{path} trong công thức trên cũng chính là giá trị độ tương đồng giữa hai thực thể dựa vào ontology và cơ sở tri thức.

b) Trọng số quan hệ giữa các thực thể dựa vào thống kê sự đồng xuất hiện trong cùng một tin

Luận án kế thừa ý tưởng của [50] về độ đo PMI, nếu hai thực thể đồng xuất hiện trong cùng một mục tin nhiều lần, thì hai thực thể này có độ tương đồng với nhau cao. Sự đồng xuất hiện của các cặp thực thể có tên trong một tập dữ liệu về tin tức bóng đá sẽ được thống kê để tính trọng số PMI. Công thức được định nghĩa như sau:

$$W_{PMI}(e_1, e_2) = log \frac{\frac{c(e_1, e_2)}{N}}{\frac{c(e_1)}{N} \times \frac{c(e_2)}{N}} \tag{4.8}$$

Trong đó:

- N là số lượng tin có sẵn trong tập dữ liệu.
- $c\,(e_1, e_2)$ là số tin trong tập dữ liệu đồng xuất hiện cả hai thực thể e_1 và e_2.

- $c\,(e_1)$ là số tin trong dữ liệu chứa thực thể e_1, và $c\,(e_2)$ là số tin trong dữ liệu chứa thực thể e_2.

Như vậy, đối với mỗi cặp thực thể bất kỳ, luận án đề xuất sử dụng hai giá trị để tính trọng số quan hệ đó là: Trọng số W_{path} (được tính dựa vào đường đi ngữ nghĩa) và trọng số W_{PMI} (được tính dựa trên thống kê sự đồng xuất hiện các cặp thực thể). Trước khi kết hợp hai trọng số này với nhau, chúng cần được chuẩn hóa như công thức (4.9):

$$w_{new} = \frac{w_{old} - MIN}{MAX - MIN} \tag{4.9}$$

Trong đó: MAX và MIN lần lượt là giá trị lớn nhất và nhỏ nhất trong chuỗi giá trị w.

Sau khi chuẩn hóa, hai giá trị W_{path} và W_{PMI} được kết hợp với nhau bằng một cặp hệ số β_{path} và β_{PMI} để tính độ tương đồng của mỗi cặp thực thể như sau:

$$Similarity_{entity}(e_1, e_2) = \frac{W_{path} \times \beta_{path} + W_{PMI} \times \beta_{PMI}}{\beta_{path} + \beta_{PMI}} \tag{4.10}$$

Ta quy ước khi $e_1 \equiv e_2$ thì giá trị $Similarity_{entity}(e_1, e_2) = 1$.

c) Phương pháp tính độ tương đồng giữa hai tin dựa vào mối quan hệ giữa các thực thể

Trước hết, ta cần định nghĩa tập các thực thể liên quan đến thực thể e là một tập chứa các thực thể có độ tương đồng với e lớn hơn 0 và được ký hiệu như sau:

$$E(e) = \{e_1, e_2, e_3, \dots, e_n\}$$

Giả sử có một tin A, tập thực thể có tên được nhận dạng trong tin A được ký hiệu như sau:

$$A = \{a_1, a_2, a_3, \dots, a_m\}$$

Với mỗi thực thể a_i trong tập A, ta xây dựng một tập các thực thể liên quan đến a_i tương ứng với $E(a_i) = \{a_{i1}, a_{i2}, a_{i3}, \dots, a_{ik}\}$. Hợp tất cả các tập $E(a_i)$ này lại (i: $1 \rightarrow m$), ta có được tập của tất cả các thực thể không nằm trong A, nhưng liên quan đến A:

$$E = \bigcup_{i:1 \rightarrow m} E(a_i)$$

Cuối cùng, ta hợp hai tập A và E lại để thu được tập A_E gọi là tập mở rộng của tin A:

$$A_E = A \cup E$$

Bước tiếp theo, giá trị xếp hạng sẽ được tính cho mỗi thực thể trong tập A_E. Mỗi giá trị xếp hạng sẽ đặc trưng cho mức độ liên quan của thực thể tương ứng với tin A. Những giá trị xếp hạng này cần thỏa mãn một số tính chất:

(1) Nếu một thực thể càng xuất hiện nhiều lần trong tin, thì giá trị xếp hạng của thực thể đó càng lớn.
(2) Nếu một thực thể liên quan đến càng nhiều thực thể xuất hiện trong tin A thì thực thể đó có giá trị xếp hạng càng lớn.
(3) Giá trị xếp hạng cũng phụ thuộc vào vị trí xuất hiện của thực thể trong tin.

Về đặc tính (3), một thực thể được xác định có thể xuất hiện ở các vị trí khác nhau trong tin như sau: tiêu đề, mô tả, chữ nổi bật (chữ đậm, tiêu đề ảnh, v.v ...) và nội dung. Vậy trọng số độ quan trọng cũng được xác định cho các vị trí của tin nêu trên lần lượt như sau:

$$W_{title} > W_{description} > W_{boldertext} > W_{content}$$

Để tính giá trị xếp hạng cho mỗi thực thể trong tập A_E, dựa trên kỹ thuật *Ranked Semantic Recommendation 2* [139], luận án cũng biểu diễn các thực thể trong ma trận, trong đó hàng đầu tiên biểu diễn các thực thể trong tập A_E và cột đầu tiên biểu diễn các thực thể trong tập A. Ma trận có dạng sau:

	$\boldsymbol{e_1}$	$\boldsymbol{e_2}$	$\boldsymbol{\dots}$	$\boldsymbol{e_q}$
$\boldsymbol{a_1}$	h_{11}	h_{12}	…	h_{1q}
$\boldsymbol{a_2}$	h_{21}	h_{22}	…	h_{2q}
$\boldsymbol{\dots}$	…	…	…	…
$\boldsymbol{a_m}$	h_{m1}	h_{m2}	…	h_{mq}

Trong ma trận trên, giá trị h_{ij} được tính như sau:

$$h_{ij} = similarity(a_i, e_j) \times WE(a_i) \tag{4.11}$$

Trong đó WE (a_i) là trọng số quan trọng của thực thể a_i trong tin. Trọng số này được tính như sau: Giả sử a_i là một thực thể xuất hiện trong tin, và $N_{title}, N_{description}, N_{boldertext}, N_{content}$ tương ứng với số lần xuất hiện của a_i trong phần tiêu đề, phần mô tả, phần chữ nổi bật hơn và phần nội dung của mục tin tức. Trọng số quan trọng của thực thể a_i được tính theo công thức dưới đây:

$$WE(a_i) = N_{title} \times W_{title} + N_{description} \times W_{description} + N_{boldertext} \times W_{boldertext} + N_{content} \times W_{content} \tag{4.12}$$

Cuối cùng, theo công thức được định nghĩa trong [139], trọng số xếp hạng của mỗi thực thể e_j trong tập A_E được tính như sau:

$$Rank(e_j) = \sum_{i=1}^{m} h_{ij} \tag{4.13}$$

Gọi V_A là một vectơ có chứa các giá trị $Rank(e_i)$ được tính ở trên. Ta chuẩn hóa các giá trị của từng phần tử trong V_A trong phạm vi [0, 1]. Công thức chuẩn hóa như sau:

$$v_i = \frac{v_i - MIN}{MAX - MIN} \tag{4.14}$$

Trong đó MAX và MIN là giá trị lớn nhất và nhỏ nhất của các phần tử trong vectơ V_A. Nếu $MAX = MIN \neq 0$ thì $v_i = 1$, với mỗi giá trị của i.

Kết quả là, sau khi thực hiện tất cả các bước trên đây, ta sẽ thu được một vectơ cho mỗi tin tức. Bước cuối cùng là tính độ tương đồng giữa hai tin bất kỳ dựa trên các vectơ của chúng.

Giả sử ta có hai tin A, B và hai vectơ V_A, V_B. Vì hai vectơ này có thể không cùng số chiều, độ tương đồng giữa hai vectơ V_A, V_B (cũng là độ tương đồng giữa hai tin A và B) được xác định như một biến thể của độ tương đồng cosin, theo công thức sau:

$$similarity_{based-entity}(A, B) = cosine(V_A, V_B) = \frac{\sum_{e_a \equiv e_b}^{e_a \in A,\, e_b \in B} v_a \times v_b}{\sqrt[2]{\sum_a^{e_a \in A} v_a^2} \times \sqrt[2]{\sum_b^{e_b \in B} v_b^2}} \tag{4.15}$$

Trong đó v_a, v_b lần lượt là các giá trị $Rank(e_a), Rank(e_b)$ trong vectơ V_A, V_B .

4.3.1.2 Độ tương đồng dựa trên loại thực thể xuất hiện trong tin

Một độc giả khi quan tâm đến một đối tượng thì nhiều khả năng cũng quan tâm đến các đối tượng cùng loại khác. Ví dụ, nếu một độc giả đang đọc tin về các đội bóng, thì người đọc thường có xu hướng muốn đọc tiếp các tin tức khác cũng nói về các đội bóng chứ không phải là tin về

cầu thủ hoặc sân vận động. Do đó, nếu hai tin có sự tương đồng về các loại thực thể, thì độ tương đồng của hai tin đó sẽ cao hơn. Hình 4.1 là một ví dụ về độ tương đồng giữa hai tin dựa vào các loại thực thể trong tin tức.

Hình 4.1 *Một ví dụ về độ tương đồng giữa hai tin dựa vào các loại thực thể trong tin tức*

Trong ontology, mỗi thực thể được định nghĩa trong cơ sở tri thức sẽ thuộc về một lớp đối tượng nào đó đã được định nghĩa. Các lớp này có thể được coi là loại của thực thể. Ví dụ, hai thực thể *Lionel Messi* và *Luis Suarez* trong cơ sở tri thức là có cùng một loại, vì chúng thuộc về lớp *FootballPlayer*. Nhưng cả hai đều không cùng loại với thực thể *Barcelona-FC*, vì thực thể này thuộc về *FootballTeam*.

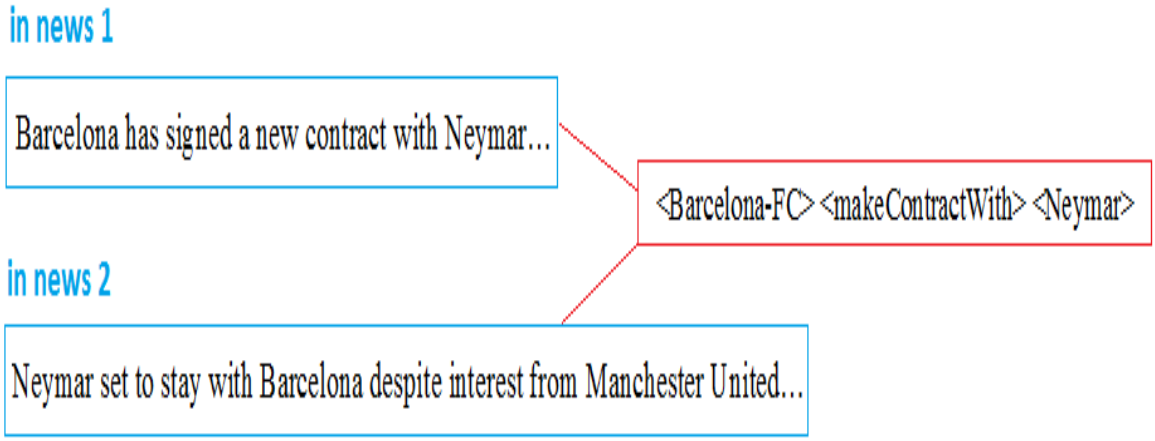
Việc thống kê các loại thực thể xuất hiện trong tin cũng tương tự như thống kê các thực thể. Hai thực thể khác nhau có thể là cùng loại. Vị trí xuất hiện của các thực thể cũng ảnh hưởng đến trọng số liên quan giữa loại thực thể với tin tương ứng. Những trọng số này sẽ được tính dựa vào tần xuất xuất hiện và vị trí xuất hiện của các thực thể thuộc loại đó. Giả sử, ta tính trọng số liên quan cho loại thực thể C đối với một tin A. Gọi c_i là các thực thể thuộc lớp C xuất hiện trong tin A, trọng số liên quan của loại thực thể C với tin A sẽ được tính theo công thức sau:

$$WC(C) = \sum WE(c_i) \tag{4.16}$$

Một vectơ cho tin được xây dựng với các phần tử là trọng số WC tương tự như xây dựng vectơ dựa trên thực thể trong mục 4.3.1.1 c) Các phần tử trong mỗi vectơ sẽ được chuẩn hóa trước khi sử dụng biến thể của công thức để tính độ tương đồng giữa các vectơ được sử dụng trong phần 4.3.1.1 c). Ký hiệu giá trị tính được này là $similartiy_{based-type}$.

4.3.1.3 Độ tương đồng dựa trên các chú thích ngữ nghĩa của tin

Các chú thích ngữ nghĩa ở đây là những bộ ba có dạng <subject> <predicate> <object>. Trong đó, subject và object là hai thực thể. Các chú thích ngữ nghĩa này cũng đóng vai trò quan trọng, vì chúng thể hiện phần nào nội dung mà tin đang nói đến. Hình 4.2 là một ví dụ về độ tương đồng giữa hai tin dựa trên các chú thích ngữ nghĩa của tin.



Hình 4.2 *Một ví dụ về độ tương đồng giữa hai tin dựa trên các chú thích ngữ nghĩa của tin*

Một tin có thể có nhiều bộ ba và một bộ ba có thể xuất hiện nhiều lần. Những bộ ba xuất hiện nhiều lần trong tin sẽ là những bộ ba quan trọng, thể hiện các nội dung chính mà tin đề cập. Hơn nữa, vị trí xuất hiện của các bộ ba này trong tin cũng thể hiện độ quan trọng của chúng. Độ quan trọng của các vị trí của tin (phần tiêu đề, phần mô tả, phần nhấn mạnh, phần nội dung) tương tự như đã trình bày ở phần trước. Hai tin mà có càng nhiều bộ ba giống nhau thì càng có độ tương đồng cao.

Với mỗi bộ ba, ta ký hiệu $N_{title}, N_{description}, N_{boldertext}, N_{content}$ lần lượt là số lần xuất hiện của bộ ba này trong phần tiêu đề, phần mô tả tin, phần nhấn mạnh và phần nội dung. Công thức được sử dụng tương tự như công thức tính trọng số quan trọng của thực thể trong mục 4.3.1.1 c), để tính trọng số quan trọng WT của mỗi bộ ba trong tin. Sau đó, các giá trị trọng số này được biểu diễn như các phần tử của một vectơ, rồi sử dụng công thức chuẩn hóa vectơ để đưa những giá trị trọng số này về đoạn [0, 1]. Để tính độ tương đồng giữa hai tin dựa vào các chú thích ngữ nghĩa, biến thể của công thức Cosine được sử dụng như trong mục 4.3.1.1 c). Để tính toán khoảng cách giữa hai vectơ, giá trị này được ký hiệu là $similarity_{based-annotation}$.

Vì vậy, 3 tham số được sử dụng để xác định độ tương đồng ngữ nghĩa giữa hai tin dựa vào các yếu tố:

- Mối liên quan giữa các thực thể có tên,
- Loại thực thể xuất hiện trong tin,
- Chú thích ngữ nghĩa của tin.

Mỗi thông số trong 3 thông số trên đều có những ý nghĩa khác nhau trong việc xác định mức độ tương đồng về ngữ nghĩa giữa hai tin. Ba thông số này lại được kết hợp với nhau để xác định được giá trị cuối cùng thể hiện mức độ tương đồng về ngữ nghĩa của hai tin. Để kết hợp 3 thông số này, một bộ 3 tham số $\theta_{entity}, \theta_{annotation}, \theta_{type}$ được sử dụng để thể hiện mức độ quan trọng của từng thông số trên. Công thức cuối cùng để tính độ tương đồng ngữ nghĩa giữa hai tin, được xác định như sau:

$$Similarity_{semantic}(A,B) = similarity_{based-entity}(A,B) \times \theta_{entity} + similarity_{based-annotation}(A,B) \times \theta_{annotation} + similarity_{based-type}(A,B) \times \theta_{type} \quad (4.17)$$

4.3.2 Độ tương đồng về nội dung

Với phương pháp gợi ý tin tức chỉ sử dụng độ tương đồng ngữ nghĩa như đề xuất ở trên, chúng ta có thể gặp phải một số vấn đề như:

- Nhận dạng không đủ hoặc nhận dạng không chính xác thực thể có tên xuất hiện trong tin tức.
- Không chú thích ngữ nghĩa được hết nội dung của tin tức.

Sự xuất hiện các hạn chế trên xảy ra là do giới hạn của thông tin trong ontology và cơ sở tri thức. Điều này rất khó tránh khỏi bởi vì việc xây dựng ontology và cơ sở tri thức phải thực hiện thủ công hoặc bán tự động nên mất rất nhiều công sức. Hơn nữa, tri thức trong thế giới thực thường xuyên thay đổi, ví dụ như các cầu thủ mới xuất hiện hoặc các cầu thủ thay đổi câu lạc bộ thi đấu, Nó gây khó khăn cho việc cập nhật kịp thời. Để khắc phục các hạn chế này, việc kết hợp tính độ tương đồng ngữ nghĩa với độ tương đồng dựa nội dung của của hai tin được đề xuất.

Mục này mô tả độ tương đồng về nội dung được tính bằng cách sử dụng trọng số TF-IDF của từ trong tin kết hợp với phép đo cosin.

Những từ có trọng số TF-IDF cao thường là những từ quan trọng, cho biết nội dung chính của tin. Vì vậy, luận án chỉ quan tâm đến những từ có trọng số TF-IDF cao. Các bước xây dựng tập các từ quan trọng của tin bao gồm:

- *Bước 1:* Loại bỏ các từ stopword. Các từ stopword là những từ không mang ý nghĩa trong việc thể hiện nội dung tin, chẳng hạn như: "a", "an", "the", v.v.
- *Bước 2*: Chuẩn hoá các từ về dạng nguyên thể. Các động từ hoặc danh từ thường tồn tại dưới nhiều dạng tùy vào ngữ cảnh, mặc dù chúng vẫn diễn tả cùng một ý nghĩa. Ví dụ: "make", "made" và "made". Vì vậy, chúng ta sẽ chuyển tất cả chúng về dạng nguyên thể.
- *Bước 3*: Tính TF-IDF cho mỗi từ trong tin (Sau khi được chuẩn hóa trong Bước 2).
- *Bước 4*: Sắp xếp và lấy ra tốp các từ có TF-IDF cao nhất dựa vào ngưỡng xác định.

Sau các bước trên, chúng ta có được một tập các từ có TF-IDF cao nhất. Tin được biểu diễn dưới dạng một vectơ có giá trị v_k là giá trị TF-IDF của các từ trong tập trên. Độ tương đồng giữa hai tin A và B có hai tập từ quan trọng S_A, S_B, và hai vectơ tương ứng V_A, V_B sẽ được tính dựa trên biến thể công thức Cosine như sau:

$$Similarity_{TF-IDF}(A,B) = \frac{\sum_{t_a \equiv t_b}^{t_a \in S_A,\, t_b \in S_B} v_a \times v_b}{\sqrt[2]{\sum_a^{t_a \in S_A} v_a^2} \times \sqrt[2]{\sum_b^{t_b \in S_B} v_b^2}} \quad (4.18)$$

Trong đó:

- t_a, t_b là các từ tương ứng trong hai bộ S_A, S_B.
- v_a, v_b là giá trị TF-IDF của từ t_a, t_b.

4.3.3 Thuật toán gợi ý tin tức với độ tương đồng kết hợp

Để kết độ tương đồng ngữ nghĩa $Similarity_{semantic}$ và độ tương đồng nội dung $Similarity_{TF-IDF}$ của hai tin, ta sử dụng cặp trọng số $\gamma_{semantic}$, $\gamma_{content}$. Công thức kết hợp được xác định như sau:

$$Similarity_{combined}(A,B) = Similarity_{semantic}(A,B) \times \gamma_{semantic} + Similarity_{TF-IDF} \times \gamma_{content} \quad (4.19)$$

Thuật toán gợi ý tin tức, được trình bày như sau:

Đầu vào: Tin mục tiêu A và tập N tin ứng viên C.

Đầu ra: Tập K tin có độ tương đồng ngữ nghĩa với A cao nhất

- *Bước 1*: Nhận dạng thực thể có tên, chú thích ngữ nghĩa cho tin A và các tin ứng viên trong tập C.
- *Bước 2*: Xây dựng tập các từ có trọng số TF-IDF cao nhất cho tin A và các tin trong tập C.
- *Bước 3*: Với mỗi tin tức C_i trong tập C, thực hiện các bước sau:
  - Bước 3.1: Tính giá trị $Similarity_{based-entity}(A, C_i)$
  - Bước 3.2: Tính giá trị $Similarity_{based-annotation}(A, C_i)$

- Bước 3.3: Tính giá trị trị $Similarity_{based-type}(A, C_i)$
- Bước 3.4: Tính giá trị $Similarity_{semantic}(A, C_i)$ dựa vào kết quả của bước 3.1, 3.2 và 3.3.
- Bước 3.5: Tính giá trị $Similarity_{TF-IDF}(A, C_i)$
- Bước 3.6: Tính giá trị $Similarity_{combined}(A, C_i)$ dựa vào các kết quả của bước 3.4 và 3.5.

- *Bước 4:* Sắp xếp các tin C_i trong tập C theo thứ tự giảm dần theo giá trị $Similarity_{combined}(A, C_i)$.
- *Bước 5*: Lấy k tin đầu danh sách đã sắp xếp ở bước 4 để gợi ý cho tin A.

Giả sử rằng n_t là số trung bình của các thẻ trong tin và n là số tin trong tập dữ liệu C. Chúng ta thấy rằng, ở bước 1, độ phức tạp của việc nhận dạng thực thể có tên và chú thích ngữ nghĩa của một tin là O ($n_c n_t$), Trong đó, n_c là tổng các lớp, thực thể và thuộc tính trong Ontology và cơ sở tri thức. Do đó, đối với n tin trong tập C và tin A, độ đo thời gian (time complexity) của bước 1 là O($nn_c n_t$). Bước 2 chuyển n+1 tin thành vectơ TF-IDF. Vì chúng ta đã tính IDF cho tất cả các thẻ trong từ điển trước khi chạy thuật toán, *độ đo thời gian* (time complexity) khi chuyển một tin thành TF-IDF bằng độ đo thời gian tính giá trị TF cho tất cả các thẻ trong tin đó, O(n_t). Do đó độ đo thời gian (complexity) của bước 2 là O (nn_t). Mặt khác, bước 3 được lặp lại n lần cho mỗi phần tử trong C. Các bước từ 3,1 đến 3,4 là phép nhân của cặp vectơ TF-IDF, do đó độ đo thời gian của mỗi lần lặp lại là O (n_t) và của bước 3 là O (nn_t). Độ đo thời gian của thuật toán sắp xếp trong bước 4 là O (nlogn). Kết quả là độ đo thời gian của thuật toán được đề xuất là O ($nn_c n_t$ + nlogn).

4.4 Cài đặt thử nghiệm và đánh giá

4.4.1 Kịch bản thử nghiệm

Mục tiêu của tiểu mục này là đánh giá và so sánh hiệu quả của 3 phương pháp gợi ý:

- Chỉ sử dụng độ tương đồng về ngữ nghĩa giữa các tin.
- Chỉ sử dụng độ tương đồng về nội dung giữa các tin.
- Kết hợp cả hai độ tương đồng trên.

Tương tự như các nghiên cứu trước, môi trường tiến hành thực nghiệm đánh giá phương pháp gợi ý tin tức luận án đề xuất là máy tính có vi xử lý Intel Core i7, CPU 2.30 GHz với RAM 8GB, hệ điều hành Microsoft Windows Server 2008. Các thuật toán được cài đặt bằng ngôn ngữ lập trình Java.

Việc đánh giá các phương pháp khác nhau được thực hiện bằng cách đo độ chính xác. Do chưa xây dựng được hệ thống online, nên trong nghiên cứu này, luận án sử dụng phương pháp đánh giá offline để đánh giá. Để đánh giá offline, một tập N = 100 tin (ký hiệu là tập A) được chọn từ một số trang web thể thao nổi tiếng như http://www.skysports.com/, http://www.espnfcasia.com/, http: //sports.yahoo.com/ và tiếp theo, các cộng tác viên được yêu cầu đánh giá rằng một tin có liên quan hay không liên quan đến một tin khác. Sau đó, ta thu được một tập dữ liệu thử nghiệm, trong đó mỗi tin A_i sẽ có tin liên quan K_{A_i} ($0 \leq K_{A_i} \leq N - 1$) và các tin tức không liên quan ($N-1-K_{Ai}$). Các phương pháp trên được thực hiện chạy riêng cho mỗi tin A_i trong tập A và cũng sinh ra đúng K_{A_i} tin có độ tương đồng cao nhất với nó (tin A_i), sau đó so sánh với K_{A_i} tin mà cộng tác viên đã xác định trong bộ dữ liệu thử nghiệm. Ví dụ, với tin A_1, cộng tác viên phát hiện 5 tin trong 99 tin còn lại liên quan đến A_1 sau đó thuật toán tự động chạy cũng sinh ra 5 tin, rồi so sánh chúng với 5 tin mà cộng tác viên đã xác định.

Ký hiệu:

- TP_{A_i} là số tin mà thuật toán gợi ý chính xác cho tin A_i .
- FP_{A_i} là số tin mà thuật toán gợi ý không chính xác cho tin A_i
- FN_{Ai} là số tin liên quan mà thuật toán không gợi ý cho tin A_i .

Độ chính xác (precision) cho một tin A_i, được xác định theo công thức sau:

$$precision(A_i) = \frac{TP_{A_i}}{TP_{A_i} + FP_{A_i}} \quad (4.20)$$

Thực hiện theo cách trên, ta có $FP_{Ai} = FN_{Ai}$, do đó $precision\ (A_i) = recall\ (A_i)$. Trong nghiên cứu này, luận án chỉ quan tâm đến $precision$ để đánh giá các phương pháp trên. Độ chính xác cuối cùng của phương pháp trên được xác định là bình quân của các độ chính xác cho toàn bộ các tin trong tập dữ liệu thử nghiệm:

$$Precision(A) = \frac{\sum_{A_i \in A} precision(A_i)}{N} \quad (4.21)$$

Một vài thông số được dùng để xác định độ quan trọng của các thành phần khi các thành phần này được kết hợp với nhau. Trong thực nghiệm này, luận án lựa chọn giá trị các tham số trên cơ sở phân tích dữ liệu về các thực thể, quan hệ, tần suất xuất hiện theo đặc thù miền ứng dụng. Ví dụ:

- Trọng số w_p của các quan hệ trong ontology để tính W_{path} được thiết lập dựa trên việc phân tích mức độ liên kết hay kết nối các thực thể của quan hệ đó trong lĩnh vực thể thao, như đã phân tích ở mục 4.3.1.1 a): $w_{managerOf} = 0.8$, $w_{playFor} = 0.6$, $w_{stadiumOf} = 0.5$,…
- $\gamma_{semantic}$ và $\gamma_{content}$ là hai tham số được dùng khi kết hợp hai phép đo độ tương đồng về ngữ nghĩa với độ tương đồng về nội dung giữa các tin. Trên quan điểm cho rằng với các tin tức thể thao mức độ ảnh hưởng của độ tương đồng nội dung là cơ sở, độ tương đồng ngữ nghĩa đóng vai trò bổ trợ, tác giả lựa chọn $\gamma_{semantic} = 1,\ \gamma_{content} = 2$.

4.4.2 Kết quả thử nghiệm và đánh giá

Sau khi chạy 3 phương pháp riêng biệt cho tập A chứa 100 tin như là kịch bản thử nghiệm đã trình bày trong mục 4.4.1, kết quả độ chính xác thu được của mỗi phương pháp thể hiện trong Bảng 4.1.

Bảng 4.1. Độ chính xác gợi ý tin tức trong các trường hợp

	Precision
Chỉ sử dụng độ tương đồng về ngữ nghĩa (*semantic-based*)	75.8 %
Chỉ sử dụng độ tương đồng về nội dung (*content-based*)	82.2 %
Kết hợp cả hai độ tương đồng (*combined*)	85.6 %

Nhận xét kết quả thử nghiệm

Bảng 4.1 chỉ ra rằng, đối với bộ dữ liệu kiểm thử A chứa 100 tin, thì phương pháp gợi ý *semantic-based* có độ chính xác không tốt bằng phương pháp gợi ý *content-based*. Trong khi đó, nếu kết hợp cả hai độ tương đồng mang lại kết quả tốt nhất. Điều này có thể được giải thích như sau:

- Khi chỉ sử dụng độ tương đồng về ngữ nghĩa (*hướng semantic-based*), chủ yếu dựa vào các thực thể xuất hiện trong tin. Do đó, trong một số trường hợp, thuật toán gợi ý đúng các tin về các thực thể liên quan nhưng chủ đề hoàn toàn khác. Đối với một số cộng tác viên, họ sẽ xem như là không liên quan.
- Theo *hướng content-based*, chủ đề của tin được gợi ý thường khá sát với tin mục tiêu. Tuy nhiên, phương pháp này không có khả năng mở rộng chủ đề. Nếu chúng ta có hai tin đều về câu lạc bộ Barcelona, trong đó tin thứ nhất nói về thi đấu của câu lạc bộ và tin thứ hai lại nói về việc chuyển nhượng cầu thủ của Câu lạc bộ, thì *hướng content-based* lại xác định hai tin này có độ tương đồng thấp.

- Khi kết hợp cả hai độ tương đồng *về nội dung và ngữ nghĩa*, thì các tin được gợi ý sẽ khắc phục được những hạn chế của mỗi độ đo riêng biệt, dẫn đến gợi ý tin hiệu quả hơn.

4.5 Kết luận chương

Chương này đã trình bày một phương pháp gợi ý tin tức dựa trên kết hợp độ tương đồng về nội dung và ngữ nghĩa của tin. Độ đo dựa vào ngữ nghĩa được tính dựa vào mối quan hệ ngữ nghĩa giữa các đối tượng. Nó cho phép việc gợi ý không chỉ dừng ở gợi ý những tin cùng chủ đề hoặc những tin xoay quanh chủ đề chính của tin mục tiêu, mà còn có khả năng suy diễn để gợi ý những tin nói về các thực thể (đối tượng) khác mà các thực thể này có quan hệ ngữ nghĩa với các thực thể trong mục tin mục tiêu. Tuy nhiên, đo độ tương đồng chủ yếu tập trung vào các thực thể mà không đề cập tới ngữ cảnh mà tin nhắc đến. Độ đo dựa vào nội dung sẽ khắc phục nhược điểm trên của độ đo ngữ nghĩa bằng cách trích xuất trong tin những từ có chỉ số TF-IDF cao nhất và những từ này thường là những từ đặc trưng cho ngữ cảnh chính được nhắc đến trong tin.

Luận án đã đánh giá và so sánh độ chính xác của phương pháp đề xuất và phương pháp gợi ý khi chỉ sử dụng riêng lẻ từng loại tương đồng. Kết quả thử nghiệm cho thấy việc kết hợp hai độ tương đồng sẽ giúp cho nâng cao hiệu quả của cả hai phương pháp, đồng thời mỗi phương pháp lại khắc phục được điểm yếu của phương pháp kia, cuối cùng làm tăng hiệu quả của việc gợi ý tin. Kết quả nghiên cứu nói trên của luận án đã được công bố trong bài báo "Semantic-Based Recommendation Method for Sport News Aggregation System" tại hội nghị quốc tế *the 2016 International Conference on Research and Practical Issues of Enterprise Information Systems* (CONFENIS 2016).

Tuy nhiên, phương pháp đề xuất vẫn còn tồn tại một số điểm hạn chế như phụ thuộc vào độ đầy đủ của cơ sở tri thức và ontology. Việc xác định các bộ trọng số sao cho việc kết hợp các độ đo đạt được hiệu quả cao nhất cũng là một vấn đề khó khăn cần giải quyết của phương pháp và đó cũng là nhiệm vụ của những nghiên cứu sau này.

KẾT LUẬN

Căn cứ vào các chương đã trình bày trong luận án, phần này tổng kết những kết quả đạt được, đồng thời đưa ra các hạn chế chưa giải quyết được, và đề xuất hướng phát triển tiếp theo.

Các kết quả đạt được của luận án

Web ngữ nghĩa là mở rộng của Web hiện tại ở đó thông tin được bổ sung ý nghĩa rõ ràng, hỗ trợ máy và con người cộng tác với nhau tốt hơn. Với dữ liệu được định nghĩa và liên kết trên Web ngữ nghĩa, máy tính có thể xử lý, chuyển đổi, lắp ráp, tái sử dụng và tích hợp chúng qua các ứng dụng khác nhau.

Thực tế chứng tỏ rằng Web ngữ nghĩa có thể thể hiện những điểm mạnh của mình khi được áp dụng vào những lĩnh vực thông tin bị giới hạn, ví dụ quản lý tri thức, phát triển những dịch vụ Web có ngữ nghĩa. Với sự hỗ trợ của Web ngữ nghĩa, thông tin mong muốn được tìm ra nhanh hơn và chính xác hơn. Web ngữ nghĩa cũng hỗ trợ tích hợp dữ liệu liên kết từ nhiều nguồn, tìm kiếm động các dữ liệu sẵn có và các nguồn dữ liệu.

Luận án tận dụng những ưu điểm vượt trội của Web ngữ nghĩa như tìm kiếm tốt hơn, tổ chức, sắp xếp, trực quan hóa một cách tự động. Luận án đã ứng dụng công nghệ Web ngữ nghĩa để xây dựng mô hình ngữ nghĩa trong hệ thống tổng hợp tin tức thể thao được đặt tên là BKSport. Đối với người dùng, hệ thống hoạt động như trang tin tức thông thường mà ở đó người dùng có thể xem tin tức tổng hợp từ một số nguồn tin cậy và được hỗ trợ tính năng tìm kiếm và gợi ý tin tức.

Các đóng góp chính của luận án như sau:

Thứ nhất, luận án đề xuất một số phương pháp sinh chú thích ngữ nghĩa cho các tin tức thể thao bằng văn bản một cách tự động.

Ý tưởng cơ bản xuyên suốt là sử dụng ontology và cơ sở tri thức để nhận dạng và xác định lớp cho các thực thể có tên. Một số kỹ thuật được luận án đề xuất để nâng cao hiệu quả của tác vụ này là phát hiện bí danh thực thể, nhận dạng các thực thể ở mức khái niệm chi tiết, cải tiến nhận dạng thực thể có tên ở dạng rút gọn, nhận dạng thực thể cùng tên khác kiểu.

Sau đó dựa trên việc xây dựng các luật trích chọn mà các thực thể có tên là một thành phần, luận án nhận dạng và sinh thành công các dạng thức ngữ nghĩa khác nhau của tin tức thể thao bao gồm ngữ nghĩa bộ ba đơn giản để diễn tả các sự kiện, ngữ nghĩa về thực thể quan trọng trong tin tức, và một số ngữ nghĩa phức tạp như tuyên bố gián tiếp, xử lý đại từ, ngữ nghĩa chuyển nhượng.

Thứ hai, luận án đề xuất phương pháp chuyển đổi câu hỏi bằng ngôn ngữ tự nhiên tiếng Anh sang truy vấn ngữ nghĩa được biểu diễn ở dạng thức SPARQL. Truy vấn này được dùng để thực hiện tìm kiếm ngữ nghĩa. Từ đó, hệ thống thực hiện được tìm kiếm sử dụng mô tơ tìm kiếm ngữ nghĩa. Luận án đã xây dựng một mô hình ngữ nghĩa để biểu diễn truy vấn SPARQL cần sinh ra. Mô hình này có khả năng diễn đạt một số dạng câu hỏi phức tạp như câu hỏi so sánh hơn, so sánh hơn nhất, câu hỏi có nhiều chủ ngữ, tân ngữ. Nội dung cốt lõi của phương pháp nằm ở việc ánh xạ các kết quả của việc phân tích cú pháp vào quá trình sinh truy vấn trung gian cũng như hoàn chỉnh câu truy vấn. Cơ sở tri thức và ontology được khai thác để nhận dạng thực thể có tên, thuộc tính, lớp. Luận án đề xuất các kỹ thuật xử lý cụ thể để xác định các thành phần định nghĩa trong mô hình ngữ nghĩa tương ứng với các dạng câu hỏi khác nhau. Kết quả thực nghiệm cho thấy phương pháp sinh được nhiều dạng câu hỏi với độ chính xác cao phù hợp với lĩnh vực thể thao.

Thứ ba, luận án đã đưa ra độ đo tương đồng giữa hai tin tức trên cơ sở kết hợp độ liên quan ngữ nghĩa và độ tương đồng nội dung. Khác với độ tương đồng nội dung được tính dựa trên phương pháp truyền thống, độ liên quan ngữ nghĩa giữa hai tin tức là sự kết hợp của các độ liên quan ngữ nghĩa giữa các thực thể, độ tương đồng về kiểu thực thể, độ tương đồng về chú thích ngữ nghĩa của hai tin. Dựa trên độ đo nói trên, luận án phát triển phương pháp gợi ý tin tức thể thao dựa trên ngữ nghĩa.

Mặc dù luận án có mục tiêu tìm ra những phương pháp mới nhằm xây dựng hệ thống tổng hợp tin tức đem lại hiệu quả và sự thân thiện người dùng trong việc truy cập thông tin trong lĩnh vực thể thao, giá trị ứng dụng của kết quả nghiên cứu đạt được không giới hạn trong lĩnh vực này. Một số giai đoạn trong các phương pháp đề xuất trên có thể được áp dụng trong các lĩnh vực khác và có thể đem lại kết quả nếu ontology và cơ sở tri thức được xây dựng tốt cho lĩnh vực mới. Cụ thể hơn, trong số các thuật toán sinh chú thích ngữ nghĩa tự động cho tin tức, thuật toán phát hiện các ngữ nghĩa bộ ba đơn giản, tuyên bố gián tiếp và các thực thể quan trọng trong tin tức không quá phụ thuộc vào những đặc thù cụ thể của miền lĩnh vực, ngoài việc sử dụng một cơ sở tri thức. Trong nghiên cứu thứ hai, bài toán chuyển đổi câu hỏi diễn đạt bằng ngôn ngữ tự nhiên sang truy vấn SPARQL phụ thuộc khá nhiều vào các kiểu câu hỏi với ngữ nghĩa đặc thù trong lĩnh vực thể thao. Tuy nhiên, với một số dạng câu hỏi về tin tức (tài liệu) liên quan một hay nhiều thực thể, quan hệ giữa hai thực thể, phương pháp đề xuất có thể chuyển đổi thành công khi chuyển sang lĩnh vực khác. Yếu tố đặc thù miền có ảnh hưởng tới độ tương đồng giữa các tin tức mà luận án đề xuất trong nghiên cứu thứ ba chủ yếu liên quan đến các trọng số xếp hạng độ quan trọng của các quan hệ ngữ nghĩa. Do đó khi áp dụng sang một lĩnh vực khác, phương pháp này hoàn toàn có khả năng áp dụng khi các trọng số này được cập nhật.

Tóm lại kết quả của luận án đã đáp ứng được mục tiêu nghiên cứu đặt ra ban đầu. Những kết quả của luận án được thể hiện trong các công trình công bố trên các tạp chí và hội thảo chuyên ngành có phản biện trong và ngoài nước, cũng như được minh họa trên hệ thống tổng hợp tin tức BKSport đã được triển khai trong thực tế.

Hướng phát triển

Luận án đã đề xuất các phương pháp về sinh chú thích ngữ nghĩa, tìm kiếm ngữ nghĩa với câu hỏi bằng ngôn ngự nhiên, và gợi ý dựa trên ngữ nghĩa. Các phương pháp đề xuất đã đạt được một số kết quả nhất định bước đầu. Với mong muốn đưa ra một giải pháp tương đối hoàn thiện cho bài toán xây dựng hệ thống tổng hợp tin tức, luận án quan tâm đến nhiều vấn đề nghiên cứu và chắc chắn còn nhiều công việc nghiên cứu cần được thực hiện trong tương lai. Dưới đây là một số hướng nghiên cứu tiếp theo của luận án.

Trong quy trình tổng thể của hệ thống, chất lượng của tin tức đầu vào có ảnh hưởng quan trọng tới hiệu quả của các bước xử lý phía sau. Luận án cần nâng cao chất lượng của bộ thu thập tin tức Crawler nhằm loại bỏ các tin tức trùng lặp và ngoài chủ đề, và sử dụng Ontology để định hướng tác vụ nói trên đang được xem xét. Đồng thời ontology và cơ sở tri thức cần được cập nhật để theo sát với những thay đổi trong thực tế của lĩnh vực thể thao. Ví dụ, cầu thủ chuyển sang CLB khác, CLB xuống hạng lên hạng.

Đối với bài toán sinh chú thích ngữ nghĩa cho tin tức thể thao, luận án mới chỉ xem xét phát hiện một số ngữ nghĩa thường gặp và giới hạn trong phạm vi chủ đề bóng đá. Nhìn chung đa phần trong số các thuật toán đề xuất có thể áp dụng ở các chủ đề khác như ten nít, bóng rổ nhưng một số ngữ nghĩa đặc biệt thuộc những chủ đề cụ thể có thể chưa được phát hiện. Do đó một hướng nghiên cứu trong tương lai là phát hiện nhiều ngữ nghĩa phức tạp hơn từ các tin tức và biểu diễn chúng với những mô hình thích hợp. Kết quả của việc sinh chú thích ngữ nghĩa phụ thuộc vào các luật (quy tắc) trích rút. Trong tương lai, những nghiên cứu tiếp theo về sinh chú thích ngữ nghĩa sẽ nhắm vào việc học các luật trích rút để tăng cường khả năng mở rộng của tiếp cận. Tác giả cùng các cộng sự cũng sẽ nhắm vào việc trích rút các ngữ nghĩa phức tạp hơn từ các tin tức thể thao và biểu diễn chúng ở mô hình thích hợp hơn như bộ bốn.

Những nghiên cứu tiếp theo của chuyển đổi câu hỏi bằng ngôn ngữ tự nhiên tiếng Anh sang truy vấn ngữ nghĩa ở dạng thức SPARQL sẽ tập trung cải thiện những trường hợp mà hệ thống hiện tại chưa xử lý đúng và hoàn thiện ontology BKSport để bao phủ được đầy đủ các khái niệm và các quan hệ có trong miền lĩnh vực thể thao, quan tâm đến "thì" của quan hệ là động từ nhằm nắm bắt ngữ nghĩa của câu truy vấn một cách chính xác hơn. Hệ thống này sẽ được tích hợp vào cổng thông tin thể thao BKSport đã được xây dựng và đang hoàn thiện, hỗ trợ cho việc tìm kiếm tin tức hiệu quả.

Đối với bài toán gợi ý tin tức, luận án cần nghiên cứu cách thức kết hợp các độ đo về tương đồng nội dung và liên quan ngữ nghĩa hợp lý hơn việc sử dụng các trọng số được lựa chọn dựa trên thực nghiệm như hiện nay. Phương pháp hiện tại mới chỉ gợi ý dựa trên nội dung và ngữ nghĩa ẩn chứa trong tin tức. Tác giả và các cộng sự dự định mô hình hóa và sử dụng profile ngữ nghĩa về người đọc để đối sánh với ngữ nghĩa của tin tức khi gợi ý. Đây là hướng nghiên cứu đi theo tính cá nhân hóa người dùng.

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ CỦA LUẬN ÁN

1. Nguyen, Q.-M. & Cao, T.-D. (2015). A Novel Approach for Automatic Extraction of Semantic Data about Football Transfer in Sport News. *International Journal of Pervasive Computing and Communications, 11*(2), 233-252. DOI:10.1108/IJPCC-03-2015-0018.
2. Nguyen, Q.-M., Nguyen, T.-T. & Cao, T.-D. (2016). Semantic-Based Recommendation Method for Sport News Aggregation System. *Proceedings of the 2016 International Conference on Research and Practical Issues of Enterprise Information Systems (CONFENIS 2016). LNBIP 268*, pp. 32-47. Vienna, Austria: Springer. DOI:10.1007/978-3-319-49944-4_3.
3. Nguyen, Q.-M., Ngo, H.-S. & Cao, T.-D. (2018). Automatic Semantic Annotation of Sport News Using Knowledge Base and Extraction Patterns. *Journal of Science & Technology Technical Universities, 128*(06/2018), 55-62. Retrieved from http://jst.hust.edu.vn/NewsFiles/119_News_So_128_up.rar.
4. Nguyen, Q.-M., Ngo, H.-S. & Cao, T.-D. (2018). Sport News Semantic Search with Natural Language Questions. *Proceedings of the 2018 European Alliance for Innovation (EAI) International Conference on Industrial Networks and Intelligent Systems (INISCOM 2018). LNICST 257*, pp. 63-73. Da Nang, Vietnam: Springer. DOI:10.1007/978-3-030-05873-9_6.

TÀI LIỆU THAM KHẢO

[1] Akamai, "Akamai Company History", 2 March 2019. [Online]. Available: https://www.akamai.com/uk/en/about/company-history.jsp. [Accessed 2 March 2019].

[2] C. Nicholson, "WORLD CUP 2014: THE DRAMA IN THE DATA", 31 July 2014. [Online]. Available: Nicholson, C. (2014, ngày 31/07). WORLD CUP 2014: THE DRAMA IN THE DATA. The Akamai Bloghttps://blogs.akamai.com/2014/07/world-cup-2014-the-drama-in-the-data.html. [Accessed 18 February 2019].

[3] M. Castillo, "Univision, ESPN Score Digital Victories During 2014 World Cup", 15 July 2014. [Online]. Available: https://www.adweek.com/digital/univision-espn-score-digital-victories-during-2014-world-cup-158929/. [Accessed 18 February 2019].

[4] N. Adie, "Sky Sports sees record digital traffic over summer", 6 September 2013. [Online]. Available: https://www.cable.co.uk/news/sky-sports-sees-record-digital-traffic-over-summer-801634665/. [Accessed 15 February 2019].

[5] E. Fisher, "ESPN Back On Top Of ComScore Sports Ranking For March", 18 April 2018. [Online]. Available: https://www.sportsbusinessdaily.com/Daily/Issues/2018/04/18/Media/Comscores.aspx. [Accessed 19 February 2019].

[6] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web", *Scientific American*, vol. 284, no. 5, pp. 34-43, May 2001.

[7] L. Ding, T. Finin, A. Joshi, R. Pan, R. Scott Cost, Y. Peng, P. Reddivari, V. Doshi and J. Sachs, "Swoogle: A Search and Metadata Engine for the Semantic Web", in *Proceedings of the thirteenth ACM Conference on Information and Knowledge Management (CIKM 2004)*, Washingtion, D.C., USA, 2004.

[8] C. Ogbuji, E. Blackstone and C. Pierce, "Case Study: A Semantic Web Content Repository for Clinical Research", October 2007. [Online]. Available: https://www.w3.org/2001/sw/sweo/public/UseCases/ClevelandClinic/. [Accessed 15 February 2019].

[9] AGRIS, "AGRIS: International Information System for the Agricultural Science and Technology", 2019. [Online]. Available: http://agris.fao.org/agris-search/index.do. [Accessed 15 February 2019].

[10] R. Klischewski, "Semantic Web for E-Government", in *EGOV 2003*, 2003.

[11] C. Clarke, "Case Study: A Linked Open Data Resource List Management Tool for Undergraduate Students", January 2009. [Online]. Available: https://www.w3.org/2001/sw/sweo/public/UseCases/Talis/. [Accessed 15 February 2019].

[12] M. Sini, G. Salokhe, C. Pardy, J. Albert, J. Keizer and S. Katz, "Ontology-based Navigation of Bibliographic Metadata: Example from the Food, Nutrition and Agriculture Journal", in *Proceedings of the International Conference on the Semantic Web and Digital Libraries (ICSD 2007)*, Bangalore, India, 2007.

[13] H. Chen, Y. Wang, H. Wang, Y. Mao, J. Tang, C. Zhou, A. Yin and Z. Wu, "Towards a Semantic Web of Relational Databases: A Practical Semantic Toolkit and an In-Use Case from Traditional Chinese Medicine", in *The 5th International Semantic Web Conference (ISWC 2006)*, Athens, GA, USA, 2006.

[14] A. Dogac, G. B. Laleci, S. Kirbas, Y. Kabak, S. S. Sinir, A. Yildiz and G. Y., "Artemis: Deploying semantically enriched Web services in the healthcare domain", *Information Systems*, vol. 31, no. 4-5, pp. 321-339, 2006.

[15] A. Ahmad, M. Mollaghasemi and L. Rabelo, "Ontologies for Supply Chain Management", 2004.

[16] Y. A. Alsultanny, "e-learning System Overview based on Semantic Web", *The Electronic Journal of e-Learning*, vol. 4, no. 2, pp. 111-118, 2006.

[17] T. Schlachter, R. Ebel, W. Geiger, A. Sawade, M. Tauber and R. Weidermann, "Environmental Information Network of Baden-Wuerttemberg-Integration of the Authorities' Environmental Information", in *18th International Conference Informatics for Environmental Protection (EnviroInfo 2004)*, Geneva, Switzerland, 2004.

[18] J. Souer, P. Honders, J. Versendaal and S. Brinkkemper, "Defining Operations and Maintenance in Web Engineering: a Framework for CMS-based Web Applications", in *The Second IEEE International Conference on Digital Information Management (ICDIM07)*, Lyon, France, 2007.

[19] N. Suradi, H. Subramaniam, M. Hassan and S. Omar, "Development of Knowledge Portal using Open Source Tools: A Case Study of FIIT, UNISEL", *WASET International Journal of Industrial and Manufacturing Engineering*, vol. 4, no. 2, pp. 94-97, 2010.

[20] F. Christ and B. Nagel, "A Reference Architecture for Semantic Content Management Systems", in *Proceedings of the 4th International Workshop on Enterprise Modelling and Information Systems Architectures (EMISA 2011)*, Hamburg, Germany, 2011.

[21] B. Heitmann, S. Kinsella, C. Hayes and S. Decker, "Implementing Semantic Web Applications: Reference Architecture and Challenges", in *Proceedings of the 5th International Workshop on Semantic Web Enabled Software Engineering (SWESE2009)*, Washington DC, USA, 2009.

[22] M. Dumontier, "Building an effective Semantic Web for health care and the life sciences", *Semantic Web*, vol. 1, pp. 131-135, 2010.

[23] E. Hyvönen, "Semantic Portals for Cultural Heritage", in *Handbook on Ontologies – Second Edition*, Berlin, Springer-Verlag Berlin Heidelberg, 2009, pp. 757-778.

[24] F. F. Ahmed and F. S. Hmed, "Dynamic Tourism Information System Using the Semantic Web", *International Journal of Computer Science and Artificial Intelligence*, vol. 3, no. 3, pp. 120-124, 2013.

[25] Ó. Corcho, A. Gómez-Pérez, A. López-Cima, V. López-García and M. C. Suárez-Figueroa, "ODESeW. Automatic Generation of Knowledge Portals for Intranets and Extranets", in *The Second International Semantic Web Conference (ISWC 2003)*, Sanibel Island, Florida, USA, 2003.

[26] M. Stollberg, H. Lausen, R. Lara, Y. Ding, S.-K. Han and D. Fensel, "Towards Semantic Web Portals", in *Proceedings of WWW2004 Workshop on Application Design, Development and Implementation Issues in the Semantic Web*, New York, NY, USA, 2004.

[27] D. Song, C. Chute and C. Tao, "Semantator: a Semi-automatic Semantic Annotation Tool for Clinical Narratives", in *The 10th International Semantic Web Conference, Poster (ISWC2011)*, Bonn, Germany, 2011.

[28] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab and M. Strintzis, "Semantic Annotation of Images and Videos for Multimedia Analysis", in *The 2nd European Semantic Web Conference (ESWC 2005)*, Heraklion, Greece, 2005.

[29] R. Schroeter, J. Hunter and A. Newman, "Annotating Relationships Between Multiple Mixed-Media Digital Objects by Extending Annotea", in *Proceedings of The Fourth European Semantic Web Conference (ESWC 2007)*, Innsbruck, Austria, 2007.

[30] H. Cunningham, "GATE, a General Architecture for Text Engineering", *Computers and the Humanities*, vol. 36, no. 2, pp. 223-254, 2002.

[31] N. Noy, N. Shah, P. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute and M. A. Musen, "BioPortal: ontologies and integrated data

resources at the click of a mouse", *Nucleic Acids Research*, vol. 37, no. suppl_2, pp. W170-W173, 2009.

[32] T. Slimani, "Semantic Annotation: The Mainstay of Semantic Web", *International Journal of Computer Applications Technology and Research*, vol. 2, no. 6, pp. 763-770, 2013.

[33] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin and J. Y. Zien, "SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation", in *Proceedings of the Twelfth International Conference on World Wide Web (WWW 2003)*, Budapest, Hungary, 2003.

[34] P. Cimiano, S. Handschuh and S. Staab, "Towards the Self-Annotating Web", in *Proceedings of the 13th International Conference on World Wide Web (WWW 2004)*, New York, New York, USA, 2004.

[35] P. Cimiano, G. Ladwig and S. Staab, "Gimme' The Context: Context-driven Automatic with C-PANKOW", in *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*, Chiba, Japan, 2005.

[36] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov and A. Kirilov, "KIM - a semantic platform for information extraction and retrieval", *Natural Language Engineering*, vol. 10, no. 3/4, pp. 375-392, 2004.

[37] P. Kogut and W. Holmes, "AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages", in *The First International Conference on Knowledge Capture (K-CAP 2001) Workshop on Knowledge Markup and Semantic Annotation*, Victoria, British Columbia, Canada, 2001.

[38] L. Sun and X. Han, "A Feature-Enriched Tree Kernel for Relation Extraction", in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, Baltimore, Maryland, USA, 2014.

[39] A. Ben Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach", *Journal of Biomedical Semantics*, vol. 2, no. Supplement 5, p. S4, 2011.

[40] A. Bhandari and S. Batra, "SEMANTIC RETRIEVAL FOR HOMONYMS USING RDF AND SPARQL", *Journal of Global Research in Computer Science*, vol. 2, no. 4, pp. 88-91, 2011.

[41] H. Wu, G. Cheng and Y. Qu, "Falcon-S: An Ontology-Based Approach to Search Objects and Images in the Soccer Domain", in *Supplemental Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, Athens, GA, USA, 2006.

[42] S. Ferré, "SQUALL: A Controlled Natural Language for Querying and Updating RDF Graphs", in *The Third International Workshop on Controlled Natural Language (CNL 2012)*, Zurich, Switzerland, 2012.

[43] A. Bernstein, E. Kaufmann and C. Kaiser, "Querying the Semantic Web with Ginseng: A Guided Input Natural Language Search Engine", in *15th Workshop on Information Technologies and and Systems (WITS 2005)*, Las Vegas, Nevada, USA, 2005.

[44] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms", in *Proceedings of the 10th International Conference on World Wide Web (WWW 2001)*, Hong Kong, Hong Kong, 2001.

[45] A. Das, M. Datar, A. Garg and S. Rajaram, "Google News Personalization: Scalable Online Collaborative Filtering", in *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*, Banff, Alberta, Canada, 2007.

[46] X. Wu, F. Xie, G. Wu and W. Ding, "Personalized News Filtering and Summarization on the Web", in *2011 23rd IEEE International Conference on Tools with Artificial Intelligence*, Boca Raton, Florida, USA, 2011.

[47] A. Elahi, R. J. Alitappeh and A. S. Rostami, "Improvement Tfidf for News Document Using Efficient Similarity", *Research Journal of Applied Sciences, Engineering and Technology*, vol. 4, no. 19, pp. 3592-3600, 2012.

[48] A. Huang, "Similarity Measures for Text Document Clustering", in *Proceedings of the 6th New Zealand Computer Science Research Student Conference (NZCSRSC 2008)*, Christchurch, New Zealand, 2008.

[49] M. Batet, D. Sánchez and A. Valls, "An ontology-based measure to compute semantic similarity in biomedicine", *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 118-125, 2011.

[50] M. Capelle, F. Hogenboom, A. Hogenboom and F. Frasincar, "Semantic News Recommendation Using WordNet and Bing Similarities", in *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC 2013)*, Coimbra, Portugal, 2013.

[51] E. Prud'hommeaux and A. Seaborne, "SPARQL Query Language for RDF W3C Recommendation 15 January 2008", 15 January 2008. [Online]. Available: https://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/. [Accessed 19 February 2019].

[52] W3C SPARQL Working Group, "SPARQL 1.1 Overview W3C Recommendation 21 March 2013", 21 March 2013. [Online]. Available: https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/. [Accessed 19 February 2019].

[53] D. L. McGuinness and F. van Harmelen, "OWL Web Ontology Language Overview W3C Recommendation 10 February 2004", 10 February 2004. [Online]. Available: https://www.w3.org/TR/2004/REC-owl-features-20040210/. [Accessed 11 February 2019].

[54] W3C OWL Working Group, "OWL2 Web Ontology Language Document Overview (Second Edition) W3C Recommendation 11 December 2012", 11 December 2012. [Online]. Available: https://www.w3.org/TR/2012/REC-owl2-overview-20121211/. [Accessed 11 February 2019].

[55] M. Kifer and H. Boley, "RIF Overview (Second Edition) W3C Working Group Note 5 February 2013", 5 February 2013. [Online]. Available: https://www.w3.org/TR/2013/NOTE-rif-overview-20130205/. [Accessed 30 January 2019].

[56] O. Lassila, F. Van Harmelen, I. Horrocks, J. A. Hendler and D. L. McGuinness, "The Semantic Web and its languages", *IEEE Intelligent Systems & their Applications*, vol. 15, no. 6, pp. 67-73, 2000.

[57] N. Shadbolt, T. Berners-Lee and W. Hall, "The Semantic Web Revisited", *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96-101, 2006.

[58] L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neuman and S. Stephens, "The Semantic Web in Action", *Scientific American*, vol. 297, no. 6, pp. 90-97, December 2007.

[59] T. Berners-Lee, W. Hall, J. A. Hendler, K. O'Hara and N. &. W. D. J. Shadbolt, "A Framework for Web Science", *Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., ShadbolFoundations and Trends in Web Science*, vol. 1, no. 1, pp. 1-130, 2006.

[60] T. H. Lê, M. P. Từ and Q. T. Huỳnh, Tác tử công nghệ phần mềm hướng tác tử, Hanoi: Nhà xuất bản khoa học kỹ thuật, 70 Trần Hưng Đạo – Hà Nội, Vietnam, 2006.

[61] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator and W. R. Swartout, "Enabling Technology for Knowledge Sharing", *AI Magazine,* vol. 12, no. 3, pp. 36-56, 1991.

[62] T. R. Gruber, "A translation approach to portable ontology specifications", *Knowledge Acquisition,* vol. 5, no. 2, pp. 199-220, 1993.

[63] N. Guarino, "Formal Ontology and Information Systems", in *Formal Ontology in Information System 1998 (FOIS'98),* Trento, Italy, 1998.

[64] B. Swartout, R. Patil, K. Knight and T. Russ, "Toward Distributed Use of Large-Scale Ontologies", *Ontological Engineering, AAAI-97 Spring Symposium Series,* pp. 138-148, 1997.

[65] R. Studer, V. R. Benjamins and D. Fensel, "Knowledge Engineering: Principles and Methods", *Data Knowledge Engineering,* vol. 25, no. 1-2, pp. 161-197, 1998.

[66] M. M. Taye, "Understanding Semantic Web and Ontologies: Theory and Applications", *Journal of Computing,* vol. 2, no. 6, 2010.

[67] L. Ding, P. Kolari, Z. Ding and S. Avancha, "Chapter 4: Using Ontologies in the Semantic Web: A Survey", in *Ontologies: A HandBook of Principles, Concepts and Applications in Information Systems,* New York, USA, Springer, 2007, pp. 79-113.

[68] I. Horrocks, "Ontologies and the Semantic Web", *Communications of the ACM,* vol. 51, no. 12, pp. 58-67, 2008.

[69] A. Singh and P. Anand, "State of Art in Ontology Development Tools", *International Journal of Advances in Computer Science and Technology,* vol. 2, no. 7, pp. 96-101, 2013.

[70] R. V. Guha and D. B. Lenat, "Cyc: A Midterm Report", *AI magazine,* vol. 11, no. 3, pp. 32-59, 1990.

[71] M. Uschold and M. King, "Towards a Methodology for Building Ontologies", in *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95,* Montreal, Quebec, Canada, 1995.

[72] M. Grüninger and M. S. Fox, "Methodology for the Design and Evaluation of Ontologies", in *Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95,* Montreal, Quebec, Canada, 1995.

[73] G. Schreiber, B. Wielinga and W. Jansweijer, "The KACTUS View on the 'O' Word", in *IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Quebec, Canada, 1995.

[74] M. Fernández, A. Gómez-Pérez and N. Juristo, "METHONTOLOGY: From Ontological Art Towards Ontological Engineering", in *Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering*, Palo Alto, California, USA, 1997.

[75] V. Psyché, O. Mendes and J. Bourdeau, "Apport de l'ingénierie ontologique aux environnements de formation à distance", *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation, ATIEF,* vol. 10, pp. 89-126, 2003.

[76] A. Gómez-Pérez, "Toward a Framework to Verify Knowledge Sharing Technology", *Expert Systems with Application,* vol. 11, no. 4, pp. 519-529, 1996.

[77] M. Uschold and M. Grüninger, "Ontologies: Principles, Methods and Applications", *Knowledge Engineering Review,* vol. 11, no. 2, pp. 93-136, 1996.

[78] T. Tudorache, C. Nyulas, N. F. Noy and M. A. Musen, "WebProtégé: a collaborative ontology editor and knowledge acquisition tool for the Web", *Semantic Web,* vol. 4, no. 1, pp. 89-99, 2013.

[79] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer and D. Wenke, "OntoEdit: Collaborative Ontology Development for the Semantic Web", in *Proceedings of the 1st International Semantic Web Conference (ISWC2002)*, Sardinia, Italia, 2002.

[80] J. C. Arpírez, O. Corcho, M. Fernández-López and A. Gómez-Pérez, "WebODE: a Scalable Workbench for Ontological Engineering", in *Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001)*, Victoria, British Columbia, Canada, 2001.

[81] R. Mizoguchi, "Tutorial on Ontological Engineering Part 2: Ontology Development, Tools and Languages", *New Generation Computing*, vol. 22, no. 1, pp. 61-96, 2004.

[82] Y. Ding and S. Foo, "Ontology research and development. Part 2 - a review of ontology mapping and evolving", *Journal of Information Science*, vol. 28, no. 5, pp. 375-388, 2002.

[83] L. Sauermann and R. Cyganiak, "Cool URIs for the Semantic Web", 3 December 2008. [Online]. Available: http://www.w3.org/TR/cooluris. [Accessed 15 February 2019].

[84] D. Brickley and L. Miller, "FOAF Vocabulary Specification 0.99", 14 January 2014. [Online]. Available: http://xmlns.com/foaf/spec/. [Accessed 16 February 2019].

[85] P. Hayes and B. McBride, "RDF Semantics W3C Recommendation 10 February 2004", 10 February 2004. [Online]. Available: http://www.w3.org/TR/2004/REC-rdf-mt-20040210/. [Accessed 16 February 2019].

[86] P. F. Patel-Schneider, P. Hayes and I. Horrocks, "OWL Web Ontology Language – Semantics and Abstract Syntax W3C Recommendation 10 February 2004", 2 February 2004. [Online]. Available: http://www.w3.org/TR/2004/REC-owl-semantics-20040210. [Accessed 16 February 2019].

[87] B. Fazzinga and T. Lukasiewicz, "Semantic search on the Web", *Semantic Web Journal*, vol. 1, pp. 89-96, 2010.

[88] A. Seaborne, "A Query Language for RDF", 9 January 2004. [Online]. Available: http://www.w3.org/Submission/RDQL. [Accessed 12 February 2019].

[89] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis and M. Scholl, "RQL: A Declarative Query Language for RDF", in *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*, Honolulu, Hawaii, USA, 2002.

[90] J. Broekstra and A. Kampman, "SeRQL: An RDF Query and Transformation Language", in *Semantic Web and Peer-to-Peer*, Berlin, Springer-Verlag Berlin Heidelberg, 2006, pp. 23-39.

[91] M. Sintek and S. Decker, "TRIPLE – A Query, Inference and Transformation Language for the Semantic Web", in *International Semantic Web Conference 2002 (ISWC 2002)*, Sardinia, Italia, 2002.

[92] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann, "DBpedia - A crystallization point for the Web of Data", *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154-165, 2009.

[93] J. Diederich and W. Balke, "FacetedDBLP-navigational access for digital libraries", *Bulletin of IEEE Technical Committee on Digital Libraries*, vol. 4, no. 1, 2008.

[94] M. Wick, "GeoNames Ontology Version 3.1", November 2012. [Online]. Available: http://www.geonames.org/ontology/documentation.html. [Accessed 12 February 2019].

[95] J. McCrae, "The Linked Open Data Cloud version 2019-01-08", 8 January 2019. [Online]. Available: https://lod-cloud.net/versions/2019-01-08/lod-cloud.png.

[96] T. Berners-Lee, "Linked Data", 27 July 2006. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html. [Accessed 12 February 2019].

[97] D. Pollard, "Knowledge integration leading to personal knowledge management", 15 June 2004. [Online]. Available:

https://barryhardy.blogs.com/theferryman/2004/06/knowledge_integ.html. [Accessed 12 February 2019].

[98] L. Stojanovic, S. Staab and R. Studer, "eLearning based on the Semantic Web", in *WebNet2001-World Conference on the WWW and Internet*, Orlando, Florida, USA, 2001.

[99] B. Abrahams, "Tourism Information Systems Integration and Utilization Within the Semantic Web (PhD Thesis)", 2006. [Online]. Available: http://vuir.vu.edu.au/1477/. [Accessed 15 February 2019].

[100] H. Lausen, Y. Ding, M. Stollberg, D. Fensel, R. Lara Hernández and S.-K. Han, "Semantic web portals: state-of-the-art survey", *Journal of Knowledge Mangement*, vol. 9, no. 5, pp. 40-49, 2005.

[101] D. Reynolds and P. Shabajee, "SWAD-Europe deliverable 12.1.5: Semantic Portals – Requirements Specification", World Wide Web Consortium (W3C), London, 2004.

[102] Z. Jrad and M. A. Aufaure, "Personalized Interfaces for a Semantic Web Portal: Tourism Information Search", in *Proceedings of 11th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, and the 17th Italian Workshop on Neural Networks (KES 2007 / WIRN 2007), Part III*, Vietri sul Mare, Italy, 2007.

[103] J. A. DeCesare, "ARKive – An Intersection of Conservation, Multimedia and Usability", *Journal of Media Literacy Education*, vol. 4, no. 2, pp. 193-195, 2012.

[104] J. Rayfield, "BBC World Cup 2010 dynamic semantic publishing", 12 July 2010. [Online]. Available: http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_world_cup_2010_dynamic_sem.html. [Accessed 16 February 2019].

[105] C. Xu, J. Wang, H. Lu and Y. Zhang, "A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video", *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 421-436, 2008.

[106] J. Rayfield, P. Wilton and S. Oliver, "BBC Sport Ontology", 17 February 2011. [Online]. Available: http://www.bbc.co.uk/ontologies/sport/2011-02-17.shtml. [Accessed 12 February 2019].

[107] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer and R. Lee, "Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections", in *The 6th European Semantic Web Conference (ESWC 2009)*, Heraklion, Crete, Greece, 2009.

[108] S. Muthu lakshmi and G. V. Uma, "Semantic Web based e-Learning System for Sports Domain", *International Journal of Computer Applications*, vol. 8, no. 14, pp. 21-25, 2010.

[109] C. Olston and M. Najork, "Web Crawling", *Foundations and Trends in Information Retrieval*, vol. 4, no. 3, pp. 175-246, 2010.

[110] R. Iswary and K. Nath, "WEB CRAWLER", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 10, pp. 4009-4012, 2013.

[111] E. Oren, K. H. Möller, S. Scerri, S. Handschuh and M. Sintek., "What are Semantic Annotations? (Technical report)", DERI , Galway, Ireland, 2006.

[112] D. P. T. Nguyen, Y. Matsuo and M. Ishizuka, "Exploiting Syntactic and Semantic Information for Relation Extraction from Wikipedia", in *IJCAI Workshop on Text-Mining & Link-Analysis (TextLink 2007)*, Hyderabad, India, 2007.

[113] Q. D. Tran and W. Kameyama, "A Proposal of Ontology-based Health Care Information Extraction System: VnHIES", in *2007 IEEE International Conference on Research, Innovation and Vision for the Future*, Hanoi, Vietnam, 2007.

[114] K. Tymoshenko and C. Giuliano, "FBK-IRST: Semantic Relation Extraction using Cyc", in *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, Uppsala, Sweden, 2010.

[115] B. Harrington and S. Clark, "ASKNet: Creating and Evaluating Large Scale Integrated Semantic Networks", in *2008 IEEE International Conference on Semantic Computing*, Santa Clara, California, USA, 2008.

[116] H. Cunningham, D. Maynard and V. Tablan, "JAPE: a Java Annotation Patterns Engine (Research Memo CS – 00 - 10)", University of Sheffield, Sheffield, South Yorkshire, England, 2000.

[117] L. Qiu, M. Y. Kan and T. S. Chua, "A Public Reference Implementation of the RAP Anaphora Resolution Algorithm", in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 2004.

[118] T. Liang and D. S. Wu, "Automatic Pronominal Anaphora Resolution in English Texts", *Computational Linguistics and Chinese Language Processing*, vol. 9, no. 1, pp. 21-40, 2004.

[119] Q.-M. Nguyen, T.-D. Cao, H.-C. Nguyen and T. Hagino, "Towards efficient sport data integration through semantic annotation", in *The Fourth International Conference on Knowledge and Systems Engineering (KSE 2012)*, Danang, Vietnam, 2012.

[120] Q.-M. Nguyen, T.-D. Cao, T.-H. Phan, H.-C. Nguyen and T. Hagino, "A Method for the Generation of Semantic Annotation from Sport News Using Ontology Based Patterns", in *Proceedings of the 7th KES Conference on Agent and Multi-Agent Systems – Technologies and Applications (KES-AMSTA 2013)*, Hue, Vietnam, 2013.

[121] A. H. Tan and C. Teo, "Learning User Profiles for Personalized Information Dissemination", in *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks*, Anchorage, Alaska, USA, 1998.

[122] J. Zhai and K. Zhou, "Semantic Retrieval for Sports Information Based on Ontology and SPARQL", in *2010 International Conference of Information Science and Management Engineering*, Xi'an, China, 2010.

[123] T.-D. Cao and Q.-M. Nguyen, "Semantic approach to travel information search and itinerary recommendation", *International Journal of Web Information System*, vol. 8, no. 3, pp. 256-277, 2012.

[124] D. L. McGuinness, "Question Answering on the Semantic Web", *IEEE Intelligent System*, vol. 19, no. 1, pp. 82-85, 2004.

[125] Q. Guo and M. Zhang, "Question Answering System Based on Ontology and Semantic Web", in *The Third International Conference on Rough Sets and Knowledge Technology (RSKT 2008)*, Chengdu, China, 2008.

[126] C. Wang, M. Xiong, Q. Zhou and Y. Yu, "PANTO: A Portable Natural Language Interface to Ontologies", in *The 4th European Semantic Web Conference (ESWC 2007)*, Innsbruck, Austria, 2007.

[127] E. Kaufmann and A. Bernstein, "How useful Are Natural Language Interfaces to the Semantic Web for Casual End-Users?", in *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference (ISWC / ASWC 2007)*, Busan, Korea, 2007.

[128] D. Damljanovic, V. Tablan and K. Bontcheva, "A text-based Query Interface to OWL ontologies", in *The 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2008.

[129] D. Damljanovic, M. Agatonovic and H. Cunningham, "Natural Language Interfaces to Ontologies: Combining Syntactic Analysis and Ontology-based Lookup through the

User Interaction", in *7th Extended Semantic Web Conference (ESWC 2010)*, Heraklion, Crete, Greece, 2010.

[130] S. Bloehdorn, P. Cimiano, A. Duke, P. Haase, J. Heizmann, I. Thurlow and J. Völker, "Ontology-Based Question Answering for Digital Libraries", in *11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*, Budapest, Hungary, 2007.

[131] V. Lopez, E. Motta and V. Uren, "Poweraqua: Fishing the Semantic Web", in *The 3rd European Semantic Web Conference (ESWC 2006)*, Budva, Montenegro, 2006.

[132] V. Lopez, M. Pasin and E. Motta, "AquaLog: An Ontology-Portable Question Answering System for the Semantic Web", in *The 2nd European Semantic Web Conference 2005 (ESWC 2005)*, Heraklion, Crete, Greece, 2005.

[133] C. Unger and P. Cimiano, "Pythia: Compositional Meaning Construction for Ontology-Based Question Answering on the Semantic Web", in *The 16th International Conference on Applications of Natural Language to Information Systems (NLDB 2011)*, Alicante, Spain, 2011.

[134] C. Unger, L. Bühmann, J. Lehmann, A.-C. Ngonga Ngomo, D. Gerber and P. Cimiano, "Template-based Question Answering over RDF Data", in *Proceedings of the 21st International Conference on World Wide Web (WWW 2012)*, Lyon, France, 2012.

[135] M. De Marneffe and C. D. Manning, "Stanford typed dependencies manual (Technical Report)", Stanford University, Stanford, California, USA, 2016.

[136] B. Mobasher, X. Jin and Y. Zhou, "Semantically Enhanced Collaborative Filtering on the Web", in *First European Web Mining Forum (EWMF 2003)*, Cavtat-Dubrovnik, Croatia, 2003.

[137] A. M. B. Abdelrahman and A. Kayed, "A Survey on Semantic Similarity Measures between Concepts in Health Domain", *American Journal of Computational Mathematics*, vol. 5, pp. 204-214, 2015.

[138] G. Salton and C. Buckley, "TERM-WEIGHTING APPROACHES IN AUTOMATIC TEXT RETRIEVAL", *Information Processing & Management*, vol. 24, no. 5, pp. 513-523, 1988.

[139] F. Frasincar, W. IJntema, F. Goossen and F. Hogenboom, "Chapter 5: A Semantic Approach for News Recommendation", in *Business Intelligence Applications and the Web: Models, Systems and Technology*, Hershey, Pennsylvania, USA, IGI Global, 2012, pp. 102-121.

[140] B. Aleman-Meza, C. Halaschek, I. B. Arpinar and A. Sheth, "Context-Aware Semantic Association Ranking", in *Proceedings of the First International Conference on Semantic Web and Databases (SWDB 2003)*, Berlin, Germany, 2003.