

إقرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Towards Enhancing Web Browsing Experience Using Deep Annotation and Semantic Visualization

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه
حيثما ورد، وإن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل لنيل درجة أو لقب علمي أو
بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

DECLARATION

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification

Student's name:

اسم الطالب/ة: حنان محمود عبد العال

Signature:

التوقيع: حنان

Date:

التاريخ: 2016 / 03 / 8

Islamic University of Gaza
Deanery of Higher Studies
Faculty of Information Technology



Towards Enhancing Web Browsing Experience

Using Deep Annotation and Semantic

Visualization

Prepared By:

Hanan M. Abedel Aal

Supervised By:

Dr. Iyad M. Alagha

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master in Information Technology

February, 2016



نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شؤون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحثة/ حنان محمود عبدالمعطي عبدالعال لنيل درجة الماجستير في كلية تكنولوجيا المعلومات برنامج تكنولوجيا المعلومات وموضوعها:

نحو تحسين تصفح الويب باستخدام الترميز العميق والتصوير الدلالي

Towards Enhancing Web Browsing Experience Using Deep Annotation and Semantic Visualization

وبعد المناقشة التي تمت اليوم السبت 13 ربيع الآخر 1437هـ، الموافق 2016/01/23م الساعة الحادية عشرة صباحاً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

.....
الأستاذ

مشرفاً و رئيساً

د. إياد محمد الأغا

.....
الأستاذ

مناقشاً داخلياً

د. أشرف يونس مغاري

.....
الأستاذ

مناقشاً خارجياً

د. سناء وفا الصايغ

وبعد المداولة أوصت اللجنة بمنح الباحثة لجنة الماجستير في كلية تكنولوجيا المعلومات / برنامج تكنولوجيا المعلومات.

واللجنة إذ تمنحها هذه الدرجة فإنها توصيها بتقوى الله و لزم طاعته وأن تسخر علمها في خدمة دينها ووطنها.

والله والتوفيق،،،

نائب الرئيس لشؤون البحث العلمي والدراسات العليا

.....
عبد الرؤوف علي المناعمة

أ.د. عبدالرؤوف علي المناعمة

To the unconditional love in my live, my guardian Angle, my Mother

And to my beloved Father, may Allah bless his soul

Acknowledgment

First of all, I thank Allah for giving me the strength and ability to complete this study, despite all the difficulties and obstacles I've faced. I also thank my precious family: sister, brothers, and the one and only, my priceless mother; for their infinite support and encourage. I also thank all my friends who stands by me all the time.

A must to be thanks to Mr. Sandro Coelho, Java Developer at Affero Lab and cooperative developer of open source annotation tool DBpedia Spotlight, and Mr. Hector Liu, a Google Summer of Code 2012 Student Developer - DBpedia for their valuable technical tips during my programing experiments.

A great thanks to Eng. Zienab Abu Hatab – Computer and Communication Engineering Graduate, Eng. Areej Abu Rezik – Computer and Communication Engineering Graduate and Eng. Tasneem Shubair – Computer Engineering Graduate, for their valuable help in testing and evaluation tasks

I also like to show my huge gratitude and grateful to my supervisor: Dr. Iyad Mohammed Alagha for sharing his guidance and experience with me during the work of this research. Without his supervision and constant help, this thesis would not have been possible

Hanan Mahmoud Abdel Aal

Abstract

The Semantic Web annotation techniques focus on associating textual content with annotations from external resources. These annotations included additional information on the annotated terms to help users, as well as machines, to better perceive the content of the text. The Semantic Web community has proposed several approaches to integrate semantic annotation into the Web browsing activity, and created what so called "Semantic Web browsers". Despite the affordances of existing Semantic Web browsers, they mostly focus on the semantic annotation process without considering effective ways to improve the user experience. This research builds on previous efforts on Semantic Web browsers, and seeks additional techniques to make the annotation process more constructive for Web browsing. We propose two extensions to the semantic annotation process: 1) Deep annotation, which aims to find more extended, correlated and indirectly observable entities even if these entities are not contained in the Web page. 2) a semantic network that visualizes the relationships between the different terms (entities) included in the Web page being browsed. We think that the proposed techniques will help the user better interpret the Web page content and utilize semantic annotations to gain broader knowledge. Our proposed annotation process was assessed by three human subjects, and results showed that 94.12% of the retrieved annotations were correct. Results also indicated that 95.44% of the terms included in the constructed semantic network was correct.

Keywords: *Semantic Web , Semantic Annotation , Deep Annotation , Semantic Network.*

Table of Contents

Acknowledgment.....	II
Abstract.....	III
Table of Contents	IV
List of Figures	VIII
List of Tables	X
Abbreviations and Acronyms.....	XI
Chapter 1: Introduction	1
1.1 Semantic Web Browsers	1
1.2 Research Questions.....	3
1.3 Statement of the problem.....	4
1.4 Objectives	4
1.4.1 Main Objective.....	4
1.4.2 Specific Objective	4
1.5 Scope and limitation of the research	5
1.6 Outline of the Thesis.....	5
1.7 Summary	5
Chapter 2: Background	6
2.1 Semantic Web	6
2.1.1 Domain Ontology	7
2.1.2 RDF and SPARQL	8
2.2 Web Services	10
2.2.1 Big web services.....	10
2.2.2 RESTful web service	11
2.3 Semantic Annotation	12
2.3.1 Manual Semantic Annotation Tools	13
2.3.2 Semi-automatic Semantic Annotation Tools	13
2.3.3 Automatic Semantic Annotation Tools	13

2.4	Linked Open Data.....	13
2.4.1	DBpedia	14
2.5	Summary	15
Chapter 3: State of the Art and Related works.....		16
3.1	Semantic Web Browsers	16
3.2	Using Linked Open Data (LOD) in Semantic Annotation.....	18
3.3	Semantic Visualization	20
3.4	Summary	22
Chapter 4: Proposed Approach.....		23
4.1	Overview	23
4.2	System Architecture	28
4.3	System Procedures.....	30
4.3.1	Server Side (RESTful Web Service)	31
4.3.1.1	Content Extraction	31
4.3.1.2	Key Terms Identification	32
A.	DBpedia Spotlight and Named Entity Recognition.....	32
B.	DBpedia Spotlight Disambiguation Confidence	34
4.3.1.3	Semantic Annotation.....	34
A.	Informative Annotation Process	35
B.	Deep Annotation Process.....	35
4.3.1.4	Annotation Builder.....	38
A.	Annotations	38
B.	Semantic Network	38
a.	Determine the Semantic Network Central Node	40
b.	Building the Semantic Network	41
4.3.2	Returning results to the Client Side.....	42
4.3.2.1	Semantic Integration.....	42
4.4	Summary	43
Chapter 5: Design and Implementation		44
5.1	Overview	44

5.2	Server Side (RESTful Web Service)	44
5.2.1	Content Extraction	44
5.2.2	Key Terms Identification	44
	(a) DBpedia Spotlight Disambiguation Confidence	45
5.2.3	Semantic Annotation	46
5.2.3.1	Informative Annotation Process	46
5.2.3.2	Deep Annotation Process.....	48
5.2.4	Semantic Network.....	52
5.3	Returning results to the Client Side	54
5.3.1	Semantic Integration	55
5.4	Summary	58
Chapter 6: Evaluation		59
6.1	Overview	59
6.2	Evaluation Objective.....	59
6.3	Evaluation Framework	59
6.3.1	Data Set	59
6.3.2	Human Subjects.....	59
6.4	Evaluation Process	60
6.4.1	Semantic Annotation Evaluation	60
6.4.1.1	Results.....	60
6.4.1.2	Discussion	64
6.4.2	Deep Annotation and Semantic Network	65
6.4.2.1	Results.....	65
6.4.2.2	Discussion	66
6.5	Summary	67
Chapter 7: Conclusion and Future Work		68
7.1.	Conclusion	68
7.2.	Future Work	68
A.	Appendix A.....	A1
A.1	Experimental Testing	A1

A.1.1 Text Extraction	A1
A.1.2 Key Terms Identification	A3
A.1.3 Semantic Annotation Accuracy	A6

List of Figures

Figure 1-1: Enrico Motta's home page viewed through Magpie	2
Figure 2-1: Berners-Lee's Semantic Web Architecture	7
Figure 2-2: A generic RDF description	8
Figure 2-3: The general form of a SPARQL query	9
Figure 2-4: Simple SPARQL query	9
Figure 2-5: Service Request-Response mechanism	10
Figure 2-6: Basic SOAP message structure	11
Figure 2-7: Snapshot of a part of the DBpedia ontology	14
Figure 3-1: browsing a collection of videos over the web semantically.....	17
Figure 3-2: A screenshot from DBpedia Mobile.....	20
Figure 3-3: A screenshot of RelClus	21
Figure 3-4: RelFinder and relationships between Kurt Gödel and Albert Einstein	22
Figure 4-1: Snapshot of the proposed system	24
Figure 4-2: Annotated web page snap shot	25
Figure 4-3: Extracted DBpedia Information for a specific term.....	25
Figure 4-4: A visual illustration of relation between key terms.....	27
Figure 4-5: Semantic graph info-box.....	28
Figure 4-6: Proposed Approach Architecture	29
Figure 4-7: System Procedures.....	30
Figure 4-8: Extract Content from web page	32
Figure 4-9: DBpedia Spotlight Default Workflow	33
Figure 4-10: Semantic Annotation Process using DBpedia LOD	35
Figure 4-11: The Proposed Deep Annotation Approach	36
Figure 4-12: A hyperlinked text to generate the semantic network	39
Figure 4-13: JSON representation results	43
Figure 5-1: Querying DBpedia to extract resource information	46
Figure 5-2: DBpedia comment and a part of DBpedia abstract for a DBpedia resource	47
Figure 5-3: SPARQL query sample to retrieve relation between two resources with path length of one.....	48
Figure 5-4: Simple SPARQL query to retrieve the relation between two resources with path length of two.....	48

Figure 5-5: Simple SPARQL query to retrieve the relation between two DBpedia resources with path length of three.....	49
Figure 5-6: A sample relation between DBpedia resource Alstom and DBpedia resource France with path length of two.....	49
Figure 5-7: A sample relation between DBpedia resource Alstom and DBpedia resource France with path length of three	49
Figure 5-8: SPARQL query to retrieve the related resources to a given DBpedia resources with path length of two	50
Figure 5-9: A non-English results for a given Sparql Query	50
Figure 5-10: JSON representation results.	51
Figure 5-11: Central node and orbits around it	53
Figure 5-12: Info-box and related word information.....	54
Figure 5-13: Annotated term details.....	55
Figure 5-14: The proposed Firefox plugin icon	56
Figure 5-15: AJAX function used to invoke the Restful web service	56
Figure 5-16: Annotated terms on a web page	58
Figure 6-1: relation caption between France and Americans	66

List of Tables

Table 2-1: Comparison between REST and SOAP web services	12
Table 6-1: Extracted Key Term with their corresponding extracted DBpedia definition and the testing human subjects.....	61
Table 6-2: The average Precision	64
Table A-1: Extracted Data Set	A1
Table A-2: Extracted Key Terms in Experiment 1	A3
Table A-3: Extracted Key Terms in Experiment 2	A4
Table A-4: Extracted Key Terms in Experiment 3	A4
Table A-5: Extracted Key Terms in Experiment 4	A5
Table A-6: Extracted Key Terms in Our proposed approach	A5
Table A-7: Information Retrieval Accuracy in Experiment 1	A6
Table A-8: Information Retrieval Accuracy in Experiment 2	A13
Table A-9: Information Retrieval Accuracy in Experiment 3	A15
Table A-10: Information Retrieval Accuracy in Experiment 4	A17
Table A-11: Information Retrieval Accuracy in Our proposed approach	A18

Abbreviations and Acronyms

CSS	Cascading Style Sheet
HTTP	Hyper Text Transfer Protocol
LOD	Linked Open Data
NER	Named Entity Recognition
NLP	Natural Language Process
RDF	Resource Description Framework
REST	REpresentational State Transfer
SA	Semantic Annotation
SOAP	Simple Object Access Protocol
SPARQL	Simple Protocol and RDF Query Language
SW	Semantic Web
SWB	Semantic Web Browser
TF	Term Frequency
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WS	Web Service
WWW	World Wide Web
XML	EXtensible Markup Language

Chapter 1

Introduction

In the last few decades, the internet users have a fast-changing requirements while they're browsing, such as how they communicate with each other, how they discover knowledge and how to retrieve information they need over the World Wide Web (WWW) easily and clearly. To meet these on-demand requirements, Semantic Web(SW) technologies has taken a major role on how to support the generation of "intelligent" documents. An intelligent document can be defined as a document that "knows about" its own content, so that automated processes and web agents can "know what to do" with it[1]. Knowledge about documents can be provided through the annotation process, which attaches metadata concerning the world around the document, e.g. the author, and often at least part of the content, e.g. keywords. The Semantic Web suggests annotating document contents using information from domain ontologies[2].

Recently, researchers have shown an increased interest in Semantic Web annotation techniques that bind some data to some other data, so it sets a relationship between the annotated data (part of contextual content) and the annotating data (the result) in a manual, semi-automatic or fully automatic process[3].

1.1 Semantic Web Browsers

Semantic Web Browser is an application that allow the naïve user to discover the Semantic Web information by linking these information with their relevant text on the web that has been browsed[4].

Although the semantic annotations are mainly used for machine processing, many efforts[5-9]have explored how to make these annotations usable and understandable by Internet users. Therefore, Semantic Web browsers (e.g. Magpie[10], SemWeb [11]and Piggy Bank[12]) have been introduced to bridge the gap between the traditional web and the Semantic Web; they present

semantic annotations as highlights, hyperlinks, comments, graphs or in any other forms that are overlaid over Web pages and that can be easily understood by native users. To illustrate the role of the Semantic Web browser, Figure 1-1 shows the screenshot of a sample Semantic Web browser called Magpie[10]. Magpie highlights some specific terms inside the web page and annotates them with explanatory information. These terms are extracted and annotated based on a predefined Ontology, which can help the user to interpret the content of the page without having to refer to other resources.

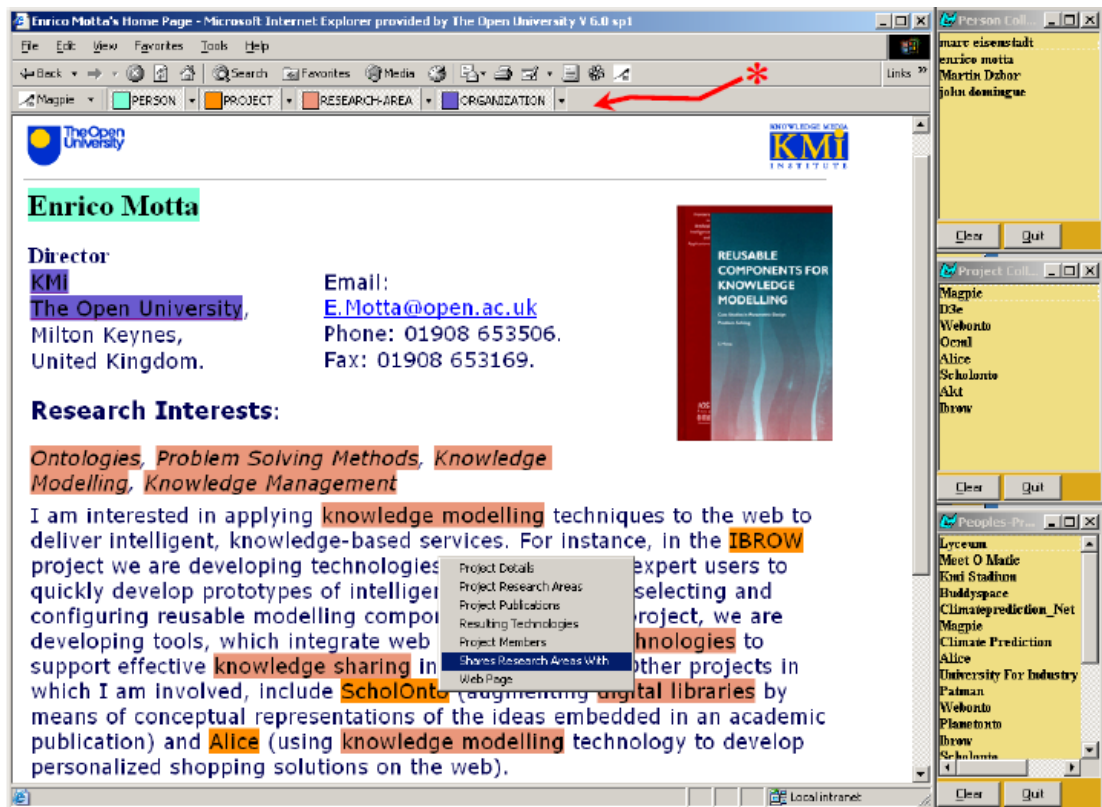


Figure 1-1: Enrico Motta's home page viewed through Magpie [10]

Despite the affordances of existing Semantic Web browsers, they mostly focus on the semantic annotation process without considering effective ways to improve the user experience. It is often necessary to think of what the user is looking for while browsing the web, and how the semantic annotation browser can be customized to address the user requirements and achieve a better user experience. This research builds on previous efforts on Semantic Web browsers, and seeks additional approaches to make the annotation process more constructive for the browsing activity. We focus on two specific shortcomings that we identified in existing Semantic Web browsers:

1. Most Semantic Web browsers focus on direct matching between Ontology/Linked Open Data (LOD) terms and the content of the page being annotated. Only terms within the document content that exactly match the Ontology/ LOD terms will be annotated. However, one term can have many synonyms which all should map to the same Ontology term. Ontologies or LOD are unlikely to cover all possible synonyms. Thus, Semantic Web annotator sometimes fails to annotate words whose synonyms, but not the exact words, are defined in the Ontology or LOD.
2. Most Semantic Web browsers primarily focus on the annotation of separate words independently without showing the relationships between the annotated terms. However, it may be useful for a user to understand the relationships between the annotated terms inside a web page. Assume, for example, that the terms “diabetes” and “high blood pressure” appear in a single web page. It may be useful for the user to determine how these terms are related to each other, even if the relationship is not directly mentioned in the text. A Semantic Web browser that reveals the different relationships between terms inside a web page can help the user gain a broad knowledge of the context.

1.2 Research Questions

Ordinary users have no technical skills to efficiently explore semantic data as quickly and easily as they need, using the existing approaches and tools such as semantic web browsers. Sometimes these tools lack the usability needed for naïve users, especially in illustrating the process of knowledge representation.

Driven by the above limitations, we propose a semantic annotation approach that provides a technical answer for our major research question, which is how to enhance the annotation process? and to solve that:

- I. We focus on how to enhance the annotation process by using deep annotation process. Unlike the usual annotation techniques, deep annotation process aims to find more extended, correlated and indirectly observable resources even if these resources are not contained in the web page as a key terms. The traditional ways focus on the directly observable entities with no ability to use hidden resources that might better

characterize the interest of the users. We think that the deep annotation will provide better understanding of the content and will generate a rich knowledge to the user. Deep annotation is discussed in the methodology section.

- II. We provide a semantic network that visualizes the relationships between the different terms (entities) included in the Web page being browsed.

1.3 Statement of the problem

This research focuses on how to provide an annotation service that can best support internet users to interpret webpages and understand the semantic relationships between annotated terms. Most existing techniques focus on annotating terms without revealing the relationships linking the annotated terms. These relationships can help users better understand the broad knowledge involved in the web page content. In addition, the provision of annotations in previous techniques is limited by the direct matching between Ontology/LOD terms and the content of the page being annotated. Some key terms may not be included in the web page but they are very important for the user to understand the context of the page.

1.4 Objectives

1.4.1 Main Objective

The aim of the research is to design an annotation service that enhances the browser activity by interpreting and annotating the content of the web page and identifying the relationships between the domain terms. Therefore, the user will be able to find all information he/she needs without having to interrupt the browsing activity to seek information from external resources.

1.4.2 Specific Objective

- Explore how to use DBpedia LOD and Wikipedia to query for and extract term descriptions.
- Design the deep annotation approach that will provide rich semantic annotations.
- Explore approaches to extract semantic relationships between the terms of the Web page.

- Investigate proper ways to integrated extracted annotations inside the web content using Java Scripts techniques.
- Integrate the proposed annotation approach as an extension to a traditional Web browser such as Firefox.
- Assess the efficiency of the annotation approach.

1.5 Scope and limitation of the research

1. The proposed annotation service will be limited to English text.
2. The proposed annotation service will be limited to Wikipedia dataset.
3. The proposed annotation service will be implemented as an extension of a commonly-used web browser such as Firefox. However, we will try to provide implementations for other Web browsers in future.

1.6 Outline of the Thesis

This thesis is organized as follows:

- Chapter 2 (Background): describes the major concepts needed for our work, such as Semantic web, Web Services, Linked Opened Data and DBpedia with a brief description for each one.
- Chapter3 (State of the art and review of related works): presents some of works and effort related to the thesis domain.
- Chapter 4 (Proposed Approach): present and discuss the proposed approach of our thesis work.
- Chapter 5 (Design and Implementation).
- Chapter 6 (Evaluation).
- Conclusion and Future Work.

1.7 Summary

In this chapter, we have introduced the thesis by describing the main concepts of semantic web browsers, how they works, and their roles on the semantic annotation process. This chapter also presents thesis research question, statement of problem, thesis objectives, and thesis outlines.

Chapter 2

Background

This chapter will introduce a brief overview of the main technical and conceptual foundation related to the thesis work. Semantic Web, Web Services, , Linked Open Data and DBpedia

2.1 Semantic Web

The Semantic Web(SW) is the major development of the traditional World Wide Web, so it is an extension of the current Web but it does not replace it. Semantic Web relies on how to deliver information with well-defined meaning to the existing web documents, and how to link between each one of them by creating a new layer of meta-data to the existing documents[2]. According to the World Wide Web consortium (W3C) the concept can be outlined as:

“The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS”.[13]

This helps to enrich users' knowledge by finding semantic relations between different data on the web content. Berners-Lee[14] proposed the structure of Semantic Web as shown in Figure 2.1.

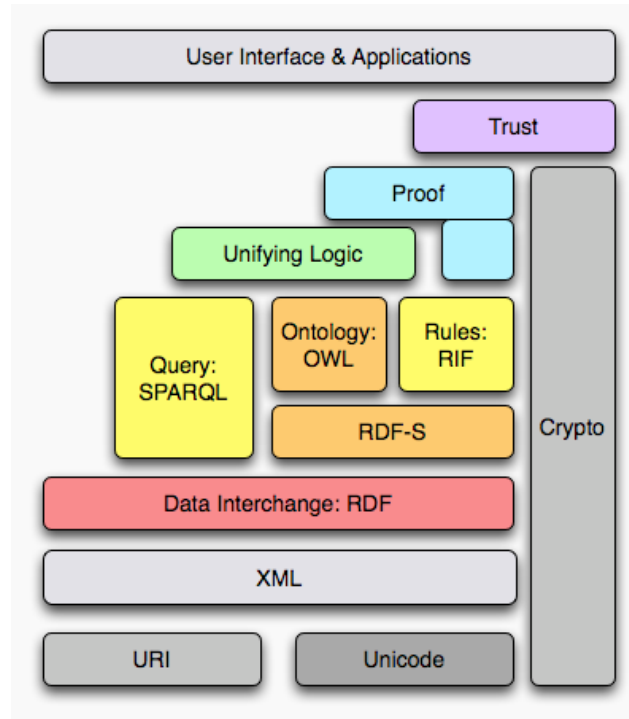


Figure 2-1: Berners-Lee's Semantic Web Architecture [15]

2.1.1 Domain Ontology

Domain Ontology is a set of related terms in a specific field or domain. It represents information related to these terms as they apply to that domain. Some of the recent studies [4-6, 16, 17] try to establish semantic annotation process based on ontological methods. Most of the proposed methods provide a simple user interaction with the web content to ensure that it was easy to understand. This kind of annotation has been written in a formal language with a well-defined semantics and referring to an ontology[18], such as Magpie[10], SemWeb[11], Haystack[4] and Piggy Bank[12].

Building a semantic annotation system depending on domain ontology brings some limitations sometimes. First, a domain ontology covers a specific domain of knowledge. Thus, annotations will be limited to the terms that belong to this specific domain alone without any others. Second, it is sometimes difficult to resolve term ambiguity by using a single ontology because of lack of information or unclear results. Third, it is hard to handle the increase in size of ontology with its huge number of classes and instances. Fourth, ontologies are often created manually, so when you have a large amount of data, it takes a lot of time to create a set of ontologies –manually - to handle it. At the same

time, the dependency between domain specific ontologies and some applications can limit the use of them to achieve any other goal in the any other purpose[19].

2.1.2 RDF and SPARQL

The Resource Description Framework (RDF) is a general framework for describing website metadata, or "information about the information" on the website. It mainly designed for machine-readable systems. RDF describes web resources as subject-predicate-object expressions which is called triple[20](See Figure 2-2). Subjects and objects are web page terms (resources) and the predicate describes the relationship between them. Each one of subjects, predicates, and objects are represented with a unique URIs, which can be abbreviated as prefixed names.

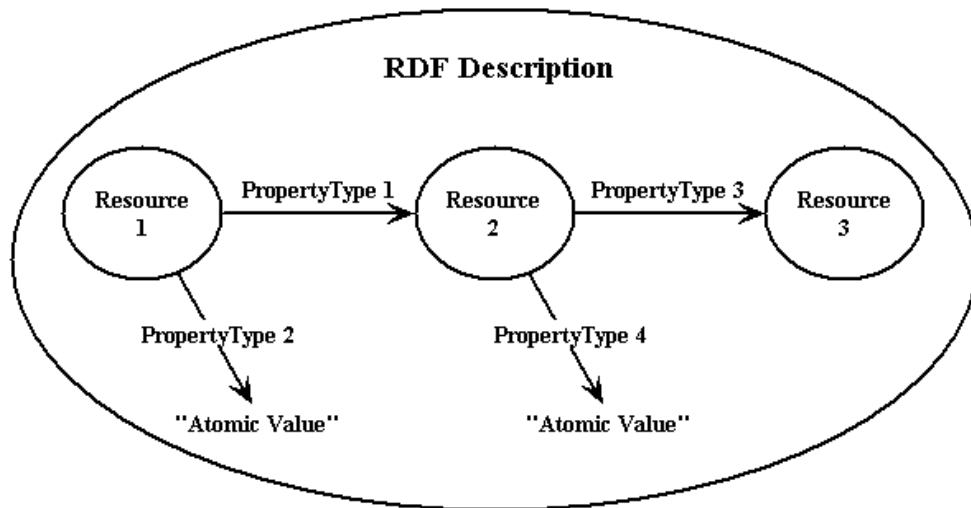


Figure 2-2: A generic RDF description[20]

SPARQL stands for Simple Protocol and RDF Query Language adopted as a W3CRecommendation[15]. It's able to query and retrieve data stored in RDF to achieve some selection criteria, and discover the relations among RDF directed Graph. It also allow user to seek data by querying and looking for unknown relationships between different resources extracted from structured data. The basic SPARQL query block consist of three parts as shown in Figure 2-3 and Figure 2-4.

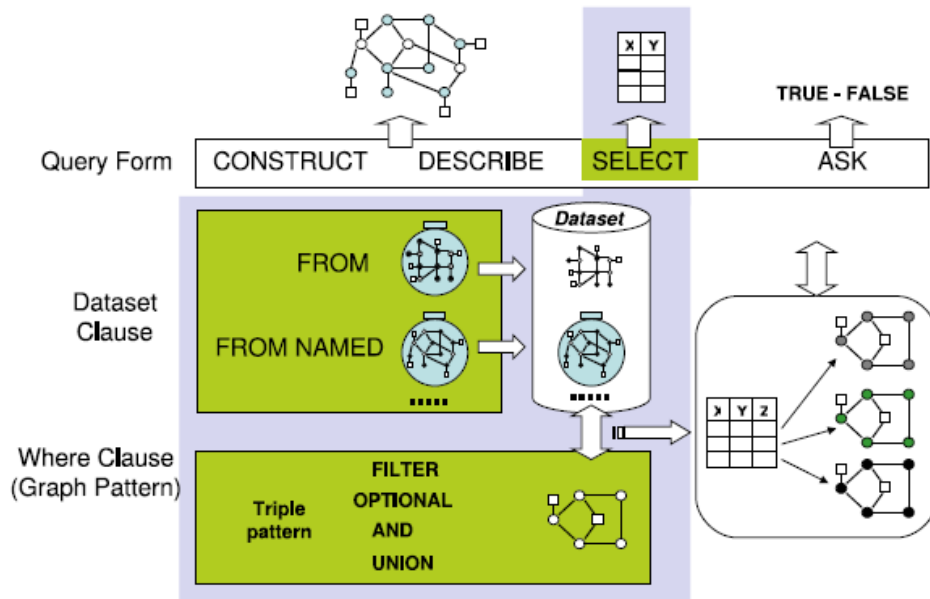


Figure 2-3: The general form of a SPARQL query[21]

1. Query Form: which determine the basic information the user wants to retrieve in a specific forms. The returned data could be a table using SELECT, a graph using DESCRIBE or CONSTRUCT, or a TRUE/FALSE answer using ASK.
2. FROM clause: which specifies the datasets sources that your query will be applied on.
3. WHERE clause: determined a set of conditions to filter the values that will be returned as a result of your query .

```

PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name
WHERE {
  ?person foaf:name ?name .
}
    
```

Figure 2-4: Simple SPARQL query

The results of SPARQL queries can be returned in different formats such as XML, JSON, CSV or RDF.

2.2 Web Services

A service is some kind of software that can be provided and accessible over network. Web services (WS) are a client-server applications that communicate between different types of applications that are running on a various platforms and frameworks over the network using Hyper Text Transfer Protocol (HTTP). As shown in Figure 2-5, the application that call the service is called a service requester, and the application response to that call and provide the data is called a service provider.

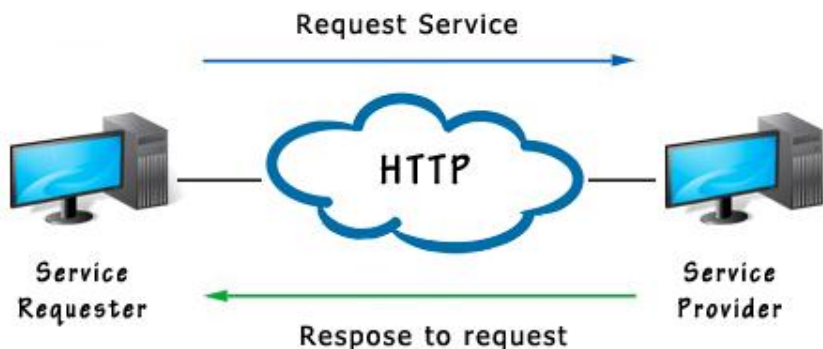


Figure 2-5: Service Request-Response mechanism

There are two major classes of Web services: “Big” web services and “RESTful” web services.

2.2.1 Big web services

Big web services is a Simple Object Access Protocol (SOAP) standard and a W3C recommendation that uses only an XML messages format for communication between applications over the internet[22]. As shown in Figure 2-6, SOAP message design must include the following elements :

- **Envelope:** A mandatory element that defines the start and the end of the message.
- **Header:** An optional element that hold every optional attributes of the message.
- **Body:** A mandatory element that contains the XML data comprising the message being sent.
- **Fault** – An optional element that provides information about errors that occur while processing the message.

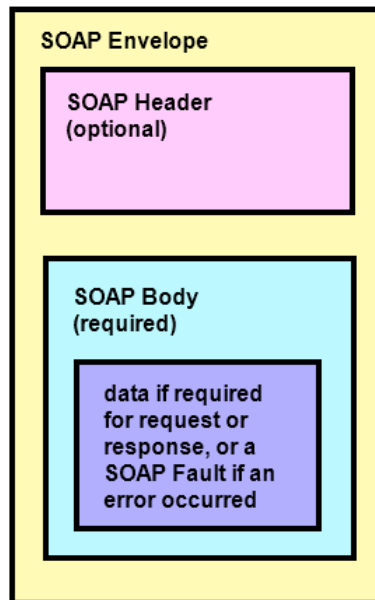


Figure 2-6: Basic SOAP message structure[23]

SOAP web services is a platform and language independent. It provides a simple communications mechanism through firewalls and proxies. Because the Soap message has to be written in XML format, and for time consuming reasons; it considers as a slow technique especially with big, long messages to be parsed.

2.2.2 RESTful web service

Representational State Transfer (RESTful) web services is an architecture style that has been described by Roy Fielding[24]. It simply sends and receives data between client and server using the following HTTP methods:

- GET (to query the current state of a resource).
- POST (to update a recourse or create one).
- PUT (to transfer the state on existing/new resource).
- DELETE (to delete an existing resource).

The different between SOAP and REST is that the REST transmit data in many different formats such as JSON, XML, or even a plain text while SOAP only use XML format. Using JSON format, for example, is better for data representation and it's fast to parse which

help browser client to perform better. Table 2-1 shows the major difference between SOAP and RESTful web services.

Table 2-1: Comparison between REST and SOAP web services

Aspect	SOAP	REST
Abbreviation	Service Oriented Architecture Protocol	Representational State Transfer
About	An XML- based message protocol	An architectural style protocol
Data Representation	XML	XML, text, JSON...etc.
HTTP Usage	Only as transport protocol (envelope)	Actions on resources applied by HTTP methods (PUT, GET, POST, DELETE)
Resource addressing	Indirect via SOAP operations	Indirect via SOAP operations
Parsing Speed and performance	Slower	Faster
Returned result	Non-human readable	Readable
Call from JavaScript	JavaScript can call SOAP, but it's difficult to implement	Easy to call from JavaScript

2.3 Semantic Annotation

Semantic annotation is the process of annotating data resources with some other data called “semantic metadata”[25]. The semantic Annotation can be useful in advance searching processes and in Information Visualization.

To achieve the annotation process, it is necessary to first determine the major terms to be annotated, by extracting useful information from a document based on some Natural Languages Processing techniques. These terms, then, have to be clarified corresponding to the same real world entity. The last phase is to associate the semantic metadata to the entities in the document through the process of annotation. These three steps are the major steps of the semantic annotation process[25].

There are several tools that used to create annotations of different web resources. These tools can be perform the semantic annotation manually, automatically or semi-automatically.

2.3.1 Manual Semantic Annotation Tools

The manual annotation transforms the existing syntactic resources into an interlinked knowledge structures that represent relevant knowledge[26]. Manual annotation tool allows users to create manual annotations in which they can be to annotate. Users also can edit text using the same tool and share it with others

2.3.2 Semi-automatic Semantic Annotation Tools

Semi-automatic annotation systems count on human involvement in the annotation process. It depends sometimes on the user's interaction to provide an initial query for providing these annotations. Vocale[9] is an example of a semi-automatic annotation tool.

2.3.3 Automatic Semantic Annotation Tools

The automatic annotation systems provide automatic proposition for annotations and sometimes, automatic annotation needs experts to achieve the best annotation a large scale. Armadillo[8] is one of these automatic annotation systems.

2.4 Linked Open Data

An alternative method for making semantic annotation is by using the Linked Open Data (LOD) to annotate the web content with its metadata. Linked open data includes semantically structured data available on the Web which has recently grown considerably. Large and important data collections, e.g. DBLP[27], CiteSeer[28] and SwissProt[29] are now available as retrievable RDF datasets. Bizer *et al*[30] defined LOD as following:

“Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external readable, defined, data sets, and can in turn be linked to from external data sets”.

The main idea behind LOD web sites is to use URIs to identify resources such as people, places, companies or even things. When these URIs invoked, a set of semantic queries will

be applied and some meaningful and useful information will be provide as a result of these queries. These results will be represents as set of Resource Description Framework (RDF).

2.4.1 DBpedia

DBpedia is an example of LOD (See Figure 2-7) which provides structured representation of Wikipedia information; it's a set of entity descriptions collection extracted from Wikipedia and represented as a linked data.

The English version of the DBpedia knowledge base describes 4.58 million things, out of which 4.22 million are classified in a consistent ontology, including 1,445,000 persons, 735,000 places (including 478,000 populated places), 411,000 creative works (including 123,000 music albums, 87,000 films and 19,000 video games), 241,000 organizations (including 58,000 companies and 49,000 educational institutions), 251,000 species and 6,000 diseases.[31]

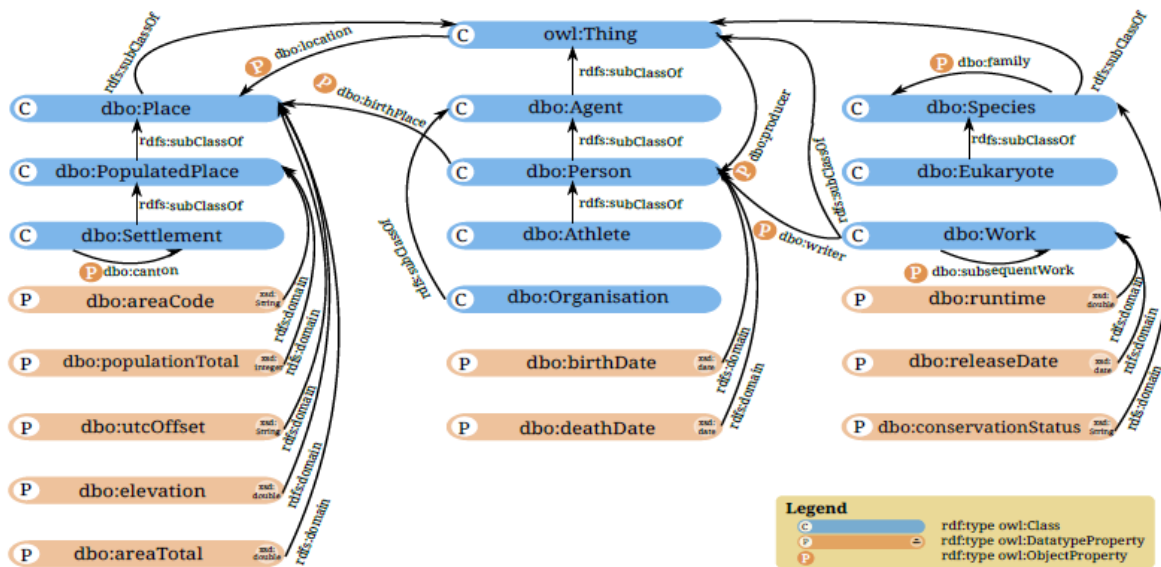


Figure 2-7: Snapshot of a part of the DBpedia ontology[32]

More than 100 languages have been covered by DBpedia such English, Chinese, German, Catalan, French... etc. Users can query DBpedia LOD using SPARQL query language and retrieve a meaningful results by exploring the relationships between different resources. The main public DBpedia SPARQL endpoint is hosted using the Virtuoso Universal Server[33].

2.5 Summary

In this Chapter we covered the basic technical and conceptual knowledge needed to understand the rest of this thesis. First we introduced the main Semantic Web concepts, especially Domain Ontology and the Semantic Web query language SPARQL. Then we presented Web Services and illustrate the major classes of Web services; “Big” web services and “RESTful” web services. Last but not least, we briefly defined the Linked Open Data and DBpedia dataset.

Chapter 3

State of the Art and Related Works

Over the past few decades, many research efforts have focused on how to use semantic annotation to enhance the web browsing process. In this chapter, we review different related works. The review focuses on semantic annotation in web browsers and how to illustrate the semantic annotation process visually. The following sections discuss semantic web browsers, using LOD in semantic annotation and semantic networks.

3.1 Semantic Web Browsers

Semantic Web browser (SWB) is a browser used for navigating the Semantic Web. it represent not only a traditional HTML web page content like any web browser, it also requests semantic information represented as RDF data about any specific resource, and illustrate to users how to navigate between there resources.

Haystack is a semantic web browser developed by Quan and Karger[5]. It allows users to create, organize, navigate, and retrieve related RDF information. It also can get metadata from many different resources, provides a flexible access to semantic web resources and presents these information to the user in a human-readable manner. The problem of this system is that it cannot resolve the problem of data integration and cleaning, which can be the covered by applying natural language processing and deep annotation to generate more relevant features.

Yoo *et al.*[34] proposed an educational tool that allow non-expert users with no technical knowledge to browse semantic web. RDF documents and HTML documents (which are hyperlinked) can be merged together as one educational environment created by teachers, and permit non-expert students to browse this environment for learning purpose. Their approach is good for naïve users (students) who have no experience on different semantic web technologies and RDF structure, but on the other hand, teachers

needs good knowledge with semantic web that can helps them to create and provide well-designed set of RDF documents and HTML documents to be hyperlinked.

Bertini *et al.* [5] studied how anyone can browse a collection of videos over the web semantically, so they develop a novel web-based tool (Figure 3-1)to provide that possibility, based on an ontology with its concepts and concepts relations. They also provide a possibility to expand the browsing from video collection to involve more related material from other sources. Our proposed approach focuses more on how to annotate textual content rather than multimedia content in any domain.

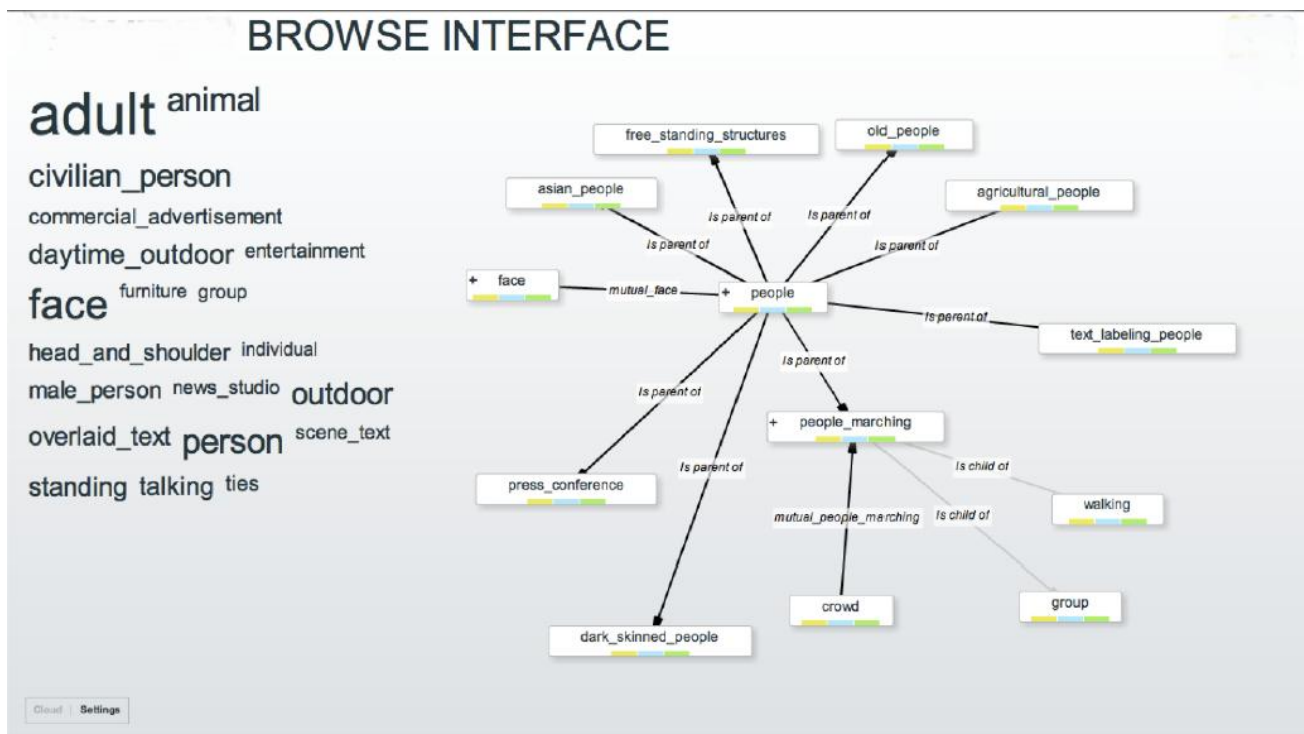


Figure 3-1: browsing a collection of videos over the web semantically[5].

Piggy Bank is another tool that was implemented as a web browser extension that developed by Huynh *et al.*[12]. This extension give the browser the ability to extract information from different websites and save them in RDF files, which give the user the possibility of browsing, searching, sorting and arrange them, no matter what is their source and types. Comparing with our approach, Piggy Bank does not allow a dynamic linking process. It allows users to create, share, retrieve and reuse semantic information on Web Source. Our proposed approach performed a knowledge extraction process to annotate the

web page contents semantically with complementary details, and dynamically links them to related resources.

Magpie is another Firefox web browser developed by Martin *et al*[10]. It provides a mechanism to highlight words that appear in web pages using a set of ontologies in a specific domain. These words has been specified using named entities recognition process to detect semantic relationships and associates entities with the specific ontological definitions. It also allows different users to collect information from the same web resource and exchange it on the basis of a common ontology. Same as Piggy Bank, Magpie does not offer a dynamic linking process unlike our approach. In addition, it could be considered as a drawback to let the user choose the ontology manually.

In the same context, many different approaches discuss how to use linked data sources such as DBpedia as a source of metadata, to identify the context of a Web according the relationships of the extracted key terms. Şah *et al.*[11] designed **SemWeb**, a Firefox web browser extension that embedded dynamic links to the web document using Semantic Web technologies and adaptive hypermedia, which will allow to create semantic hyperlink from the linked data.

Our approach builds on the previous efforts and extends them by introducing the semantic graph which shows the relationships between the domain terms. The graph enables the user to explore the domain in detail and understand how different domain concepts are semantically related. In addition, the graph can show domain terms that are highly related to the domain of interest even if the terms are not included in the web page being browsed. So it is not limited to the annotation of the web page content. Furthermore, our approach differs from Magpie and Piggybank by using DBpedia as background knowledge rather than relying on specific ontologies.

3.2 Using Linked Open Data (LOD) in Semantic Annotation

Linked Open Data helps users to publish structured data and links them between different systems as mentioned previously in Chapter 2. DBpedia is one of LOD that is used in different semantic annotation processes.

Several researchers have used DBpedia as background knowledge for their applications. Schumacher and Ponzetto[35] provided approaches for clustering searching process results based on their topical similarity, and then mapped them to DBpedia using some existing techniques. Their system takes as input a set of web search results, groups them together according to some topical standards, and provide a set of clustering as output.

Furthermore, Lama *et al.*[36] present a novel approach to extract a set of relevant terms from Learning Ontologies. These terms are identified using natural language processing techniques, and have been annotated semantically using the DBpedia Spotlight web service with relevant contextual information that has been extracted from the DBpedia using depth-limited search through the DBpedia graph.

Lukovnikov *et al.*[37] attempted to model user interests over Twitter using DBpedia entities that have been mentioned in his tweets. They relied on deep enrichment method to generate as many entities as possible to be used as extra features for classification process, which will expand entities utilizing and information collecting. This research is the start point of our work. We focus on the deep relation between any two DBpedia resources in a way that can help the naïve user to gather and understand as much as possible from the web page he/she is browsing.

The British Broadcasting Corporation (BBC) uses DBpedia Linked Data to annotated BBC different services that have been divided by domain such as food, music, news...etc. and link different services content about the same topic with different domains together through DBpedia data. They develop a new service that supports their Radio stations, TV channels and programmes brands. This can help BBC to become more consistent and helpful service by providing contextual, semantic links connecting content across different domains, so the user will not face any difficulty to find everything the BBC has published about any given subject[38].

DBpedia Mobile[39] is another application that use DBpedia in a semantic annotation process, but for mobile environments that is accessed using a mobile phone's web browser(see Figure 3-2). It allows users to access information from DBpedia about the

physical surrounding locations based on their device's GPS signal. As a result, DBpedia Mobile provides a map view annotated with related DBpedia entities.



Figure 3-2: A screenshot from DBpedia Mobile[39]

We use DBpedia dataset and DBpedia spotlight web service[40] to identify the key terms in a plain text, annotate these terms with DBpedia to extract the appropriate definitions to each one of them.

3.3 Semantic Visualization

Semantic web is considered as a large semantic network. Semantic network is a knowledge graphical representation that consist of nodes and arcs, each node represents a resource and each arch represents the relationship between different nodes.

To illustrate the semantic annotation process visually, Zhang *et al.*[41] also used DBpedia dataset to present a new clustering-based exploratory relationship search engine that groups the result of the search process automatically into a hierarchy with meaningful labels that visualized as a collapsible/expandable tree (Figure 3-3). The semantic graph which is called **RelClus**, helps the user to get and find the information he needs, and even more, he can get more information about a new key terms that are related to the extracted entities which have a relationships between each other.

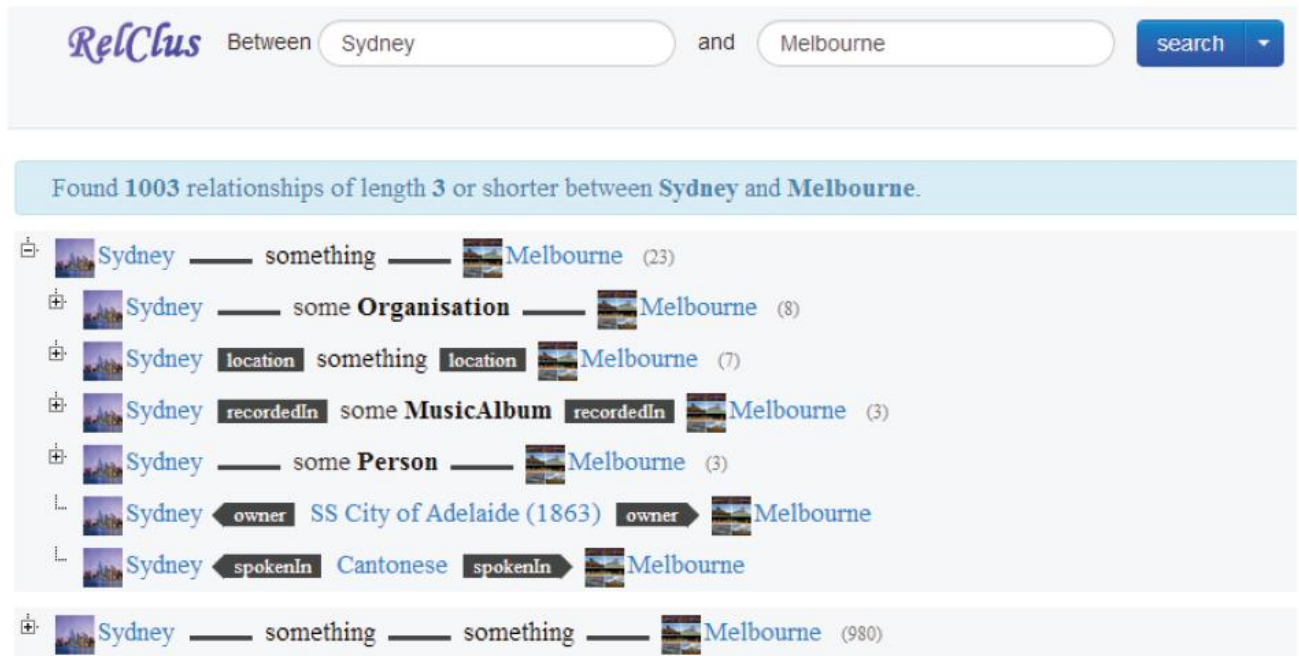


Figure 3-3: A screenshot of RelClus [41]

In like manner, Mirizzi *et al.*[7] evolve a novel approach for exploratory Knowledge search for DBpedia via new computed association between DBpedia nodes. The proposed tool allows the user to explore visually what he probably did not know to exist during the Knowledge discovery process.

Heim *et al.*[42] proposed **RelFinder**, which also known as Relationship Finder, which is a web based application that automatically discover relationships between any two specific resources in DBpedia LOD and represent them as a graph. They used properties in semantically annotated data to automatically find relationships between any pair of user-defined objects and visualize them as shown in Figure 3-4

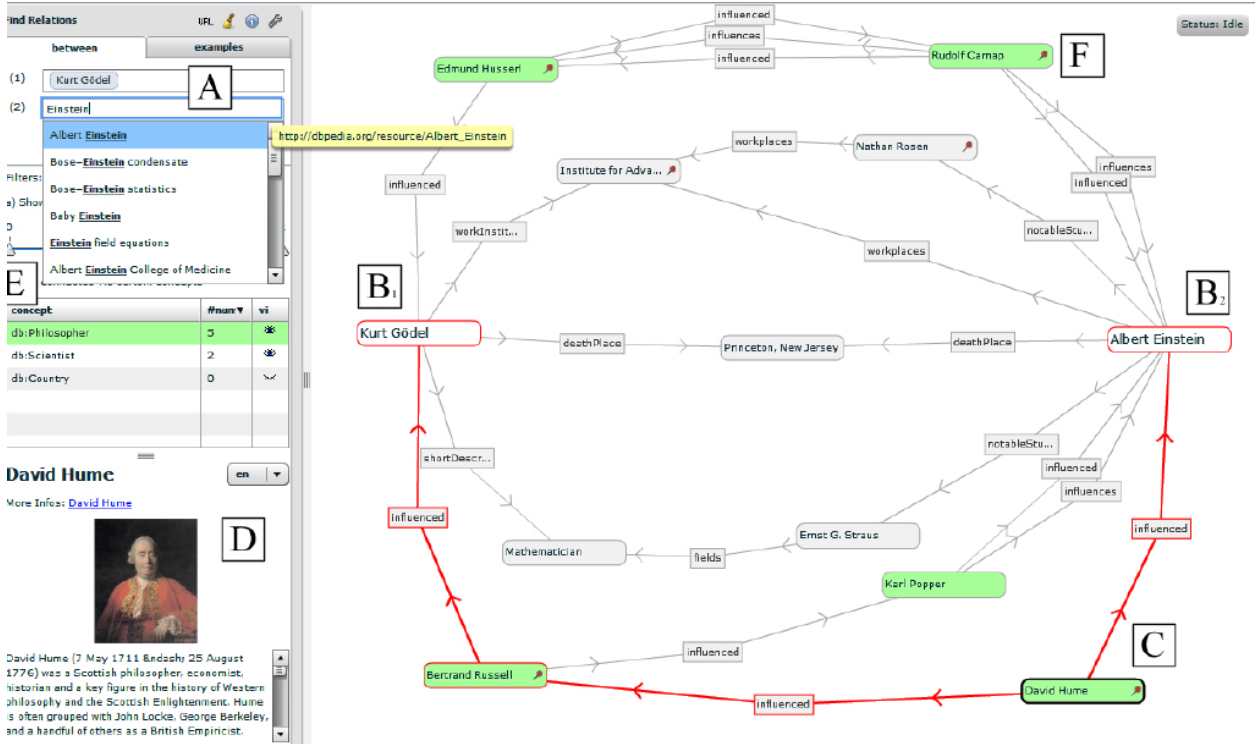


Figure 3-4: RelFinder and relationships between Kurt Gödel and Albert Einstein[42]

While the above efforts used semantic networks as standalone applications, we aim to integrate the semantic network in the browsing activity and use it as part of our proposed annotation service. We aim to find out how the semantic network can help the user perceive the content of the web page.

3.4 Summary

In this chapter, we have reviewed a set of approaches that is related to ours. The related work categorized as a semantic browser, using LOD in semantic annotation and semantic network.

Chapter 4

Proposed Approach

4.1 Overview

This chapter will present and discuss the proposed approach of our work. The research aims to explore how to annotate the key terms on the web page content semantically, and how to enhance the annotation process using deep annotation. The relationships between these key terms can be illustrated by modeling a semantic network in order to enhance web browsing and knowledge representation process. The proposed approach has the following design principles:

1. Annotations are extracted from DBpedia on the fly and are attached to the web page being browsed by the user.
2. The proposed service does not only annotate DBpedia terms mentioned in the Web page, but it also shows the semantic links connecting these terms. The aim is to enable the user to explore the topics in depth, and to understand how different terms are semantically related.
3. The proposed service should not modify the user-friendly way adopted by users while navigating the Web. Users should keep using their favorite browser while being able to annotate the page content. This has been achieved by implementing our annotation service as a plug-in to the browser rather than a standalone application.

The proposed approach contributes towards bridging the gap between the semantic web and traditional web. This can be achieved by enriching the content of non-structured web pages with complementary knowledge and associations extracted from DBpedia. Figure 4-1 shows a snapshot of the system. The figure shows the different windows associated with the system. Each window will be explained in what follows.

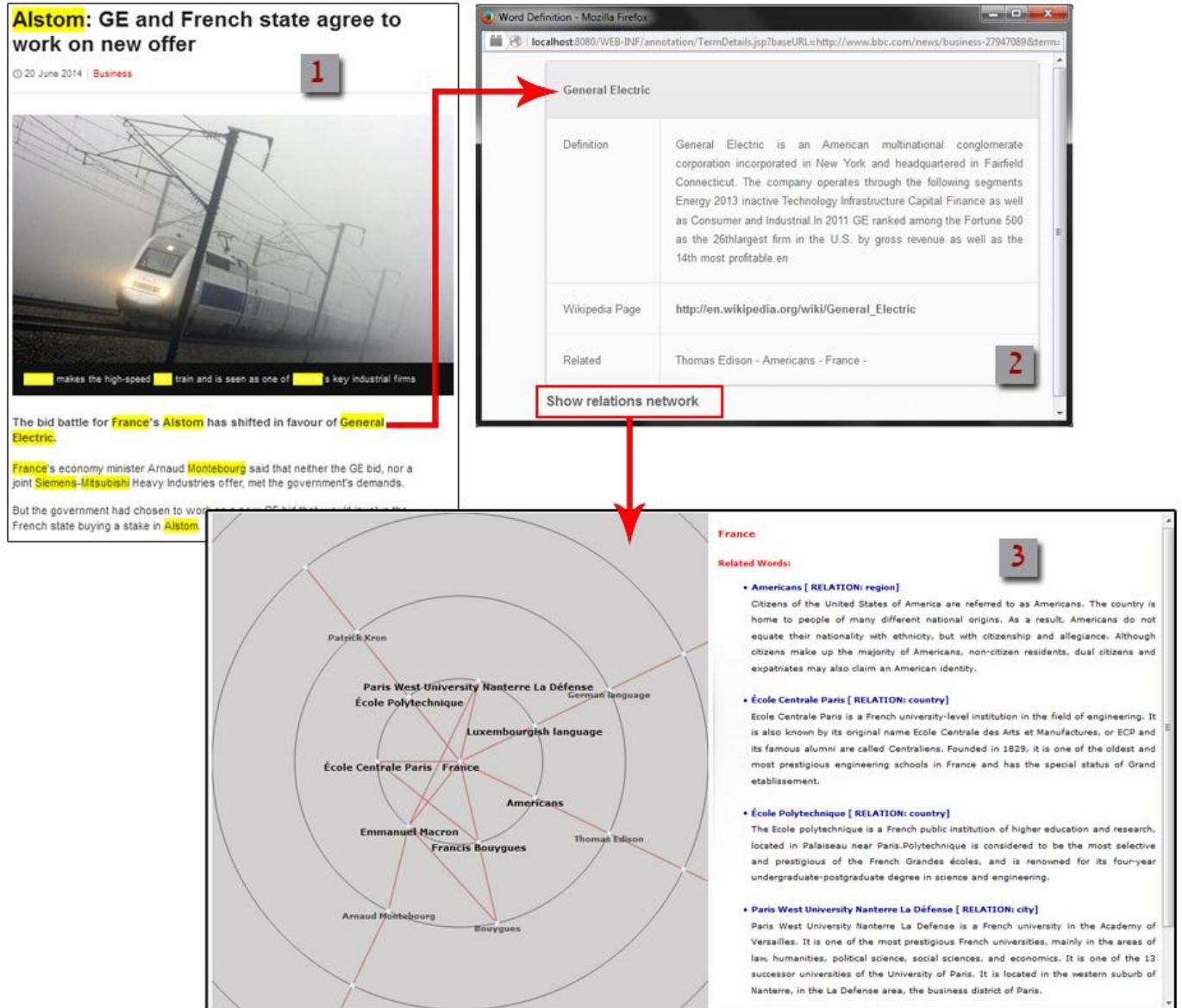


Figure 4-1: Snapshot of the proposed system

Figure 4-2 shows the web page after being annotated. Each annotated term is represented as a colored hyperlink.

Alstom: GE and French state agree to work on new offer

© 20 June 2014 | Business



The bid battle for France's Alstom has shifted in favour of General Electric.

France's economy minister Arnaud Montebourg said that neither the GE bid, nor a joint Siemens-Mitsubishi Heavy Industries offer, met the government's demands.

But the government had chosen to work on a new GE bid that would involve the French state buying a stake in Alstom.

Figure 4-2: Annotated web page snap shot

Clicking on any colored term will open a new pop-up window showing its properties as extracted from DBpedia as shown in Figure 4-3.

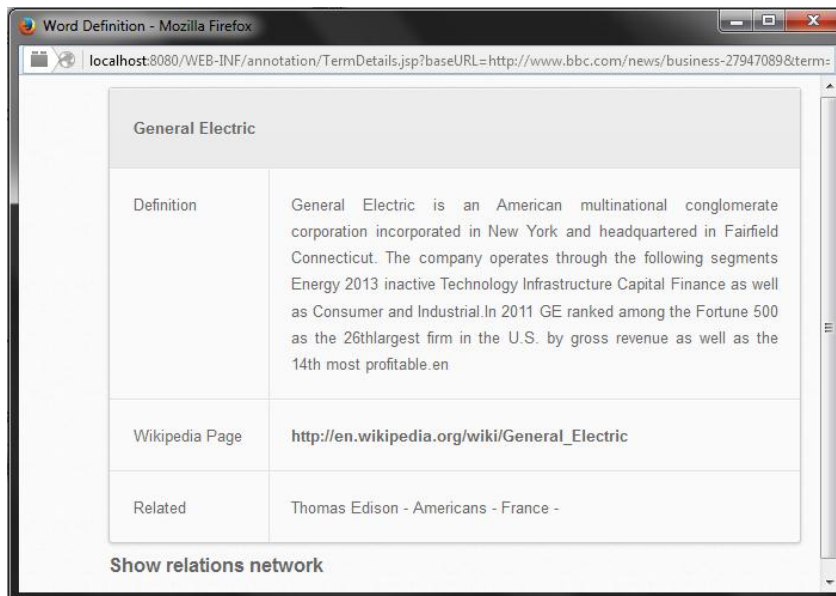


Figure 4-3: Extracted DBpedia Information for a specific term

In addition, the user can display a semantic graph showing the relations between topics in the web page. The graph also includes terms that are not mentioned in the page, but are highly related to its content. The terms that has the maximum number of connections to other terms is shown in the center of the graph. For example, the term "France" is positioned in the center of the graph in Figure 4-4. Other terms are positioned in the surrounding orbits according to the length of the semantic graph in DBpedia. Most related terms are positioned closer than less related terms.

The semantic graph is interactive. This means that clicking on any term in the graph, it will cause the graph to be rebuilt so that the newly selected term will be positioned in the center. Other terms will be reorganized around the new central term according to their relations.

The semantic graph consist of two parts: the first part shows the visual illustration of the relations between key terms (DBpedia mentions), as shown in Figure 4-4.



Figure 4-4: A visual illustration of relation between key terms

The second part is an info-box that shows basic information about every key term illustrated in the semantic network. For each term in the semantic graph, the info-box shows information about its related terms, including each related term definition, and the relation type between terms, as shown in Figure 4-5.

France

Related Words:

- **Americans [RELATION: region]**

Citizens of the United States of America are referred to as Americans. The country is home to people of many different national origins. As a result, Americans do not equate their nationality with ethnicity, but with citizenship and allegiance. Although citizens make up the majority of Americans, non-citizen residents, dual citizens and expatriates may also claim an American identity.

- **École Centrale Paris [RELATION: country]**

Ecole Centrale Paris is a French university-level institution in the field of engineering. It is also known by its original name Ecole Centrale des Arts et Manufactures, or ECP and its famous alumni are called Centraliens. Founded in 1829, it is one of the oldest and most prestigious engineering schools in France and has the special status of Grand établissement.

- **École Polytechnique [RELATION: country]**

The Ecole polytechnique is a French public institution of higher education and research, located in Palaiseau near Paris. Polytechnique is considered to be the most selective and prestigious of the French Grandes écoles, and is renowned for its four-year undergraduate-postgraduate degree in science and engineering.

- **Paris West University Nanterre La Défense [RELATION: city]**

Paris West University Nanterre La Defense is a French university in the Academy of Versailles. It is one of the most prestigious French universities, mainly in the areas of law, humanities, political science, social sciences, and economics. It is one of the 13 successor universities of the University of Paris. It is located in the western suburb of Nanterre, in the La Defense area, the business district of Paris.

Figure 4-5: Semantic graph info-box

4.2 System Architecture

As shown in Figure 4-6 , our proposed system consists of two parts: a Server Side and a Client Side. The server annotates the Webpage by matching the text of the page with the DBpedia terms. The server then creates a new copy of the Web page, and associates each DBpedia term mentioned in the text with information extracted from DBpedia. The server part also builds the semantic map that shows how different DBpedia terms are semantically associated. The annotated page as well as the generated semantic graph will be returned to the client side to be presented to the user. Note that the web pages are

annotated on-the-fly, i.e. the user does not have to wait for a long time to get response from the server.

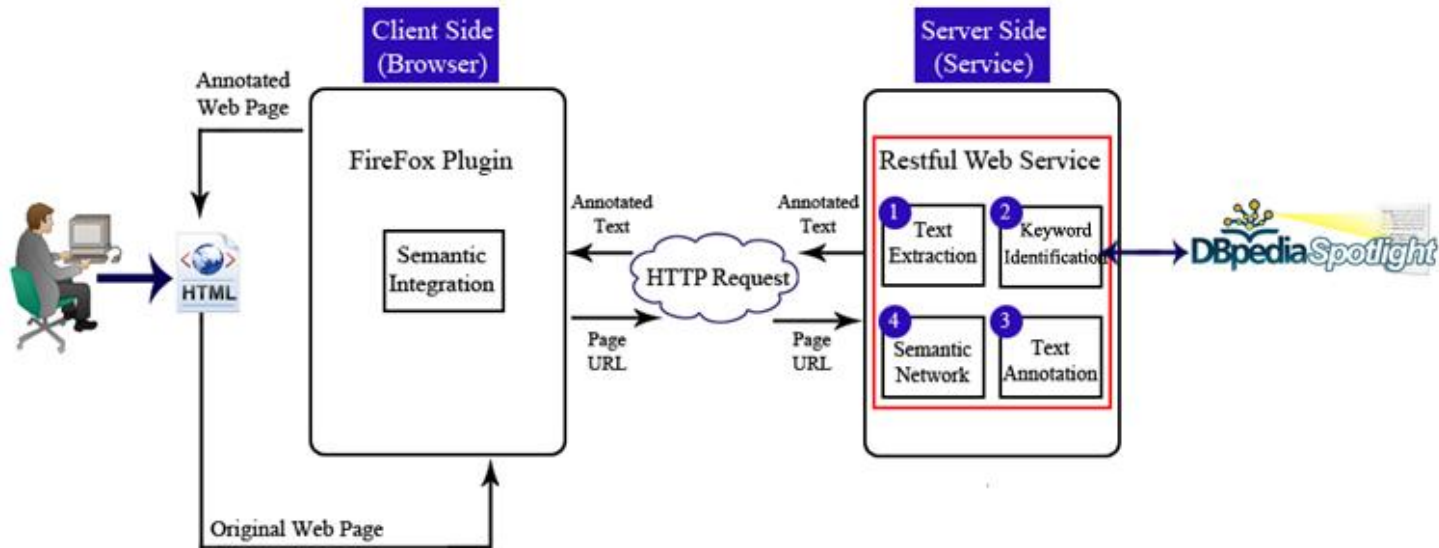


Figure 4-6: Proposed Approach Architecture

On the server side, we built a RESTful web service that performs four basic steps:

1. First step: Text Extraction; which will extract the main content of the web page with a given URL.
2. Second step: Key terms Identification; which includes determining the key terms on the extracted text using DBpedia spotlight[40].
3. Third step: Annotation Process; Once the key terms are determined, DBpedia LOD will be queried to annotate them with their definition using SPARQL query language[15] and Apache Jena[43].
4. Fourth Step: Semantic Integration; the results of annotation process will be attached and linked to the matched terms within the Web page.

The client part is implemented as a plug-in to the web browser. The plug-in shows a button that when pressed, the content of the web page being browsed will be annotated with DBpedia URIs. The client side does not involve intensive processing: it just sends the URL of the page to the server side where the annotation process takes place. All intensive tasks such as text processing, extraction of DBpedia terms and the generation of the

semantic graph are executed on the server side. The client side is only responsible for sending requests to the server, and then presenting the results to the end user.

The intention of designing the client part as a plug-in to the browser is to allow the user to exploit the annotation service while using his/her favorite web browser. Migrating the core functionality of the system to the server side is to allow for centralized maintenance and modification of the system. In addition, the lightweight processing performed by the client facilitates the development of plug-ins for different Web browsers. In our prototype system, the plug-in was built as an add-on to the Firefox web browser. Firefox was chosen due to its familiarity among internet users.

The communication between the client part, i.e. the Firefox add-on, and the server part is done through a Restful Web service. Thus, annotation requests are sent through HTTP protocol.

4.3 System Procedures

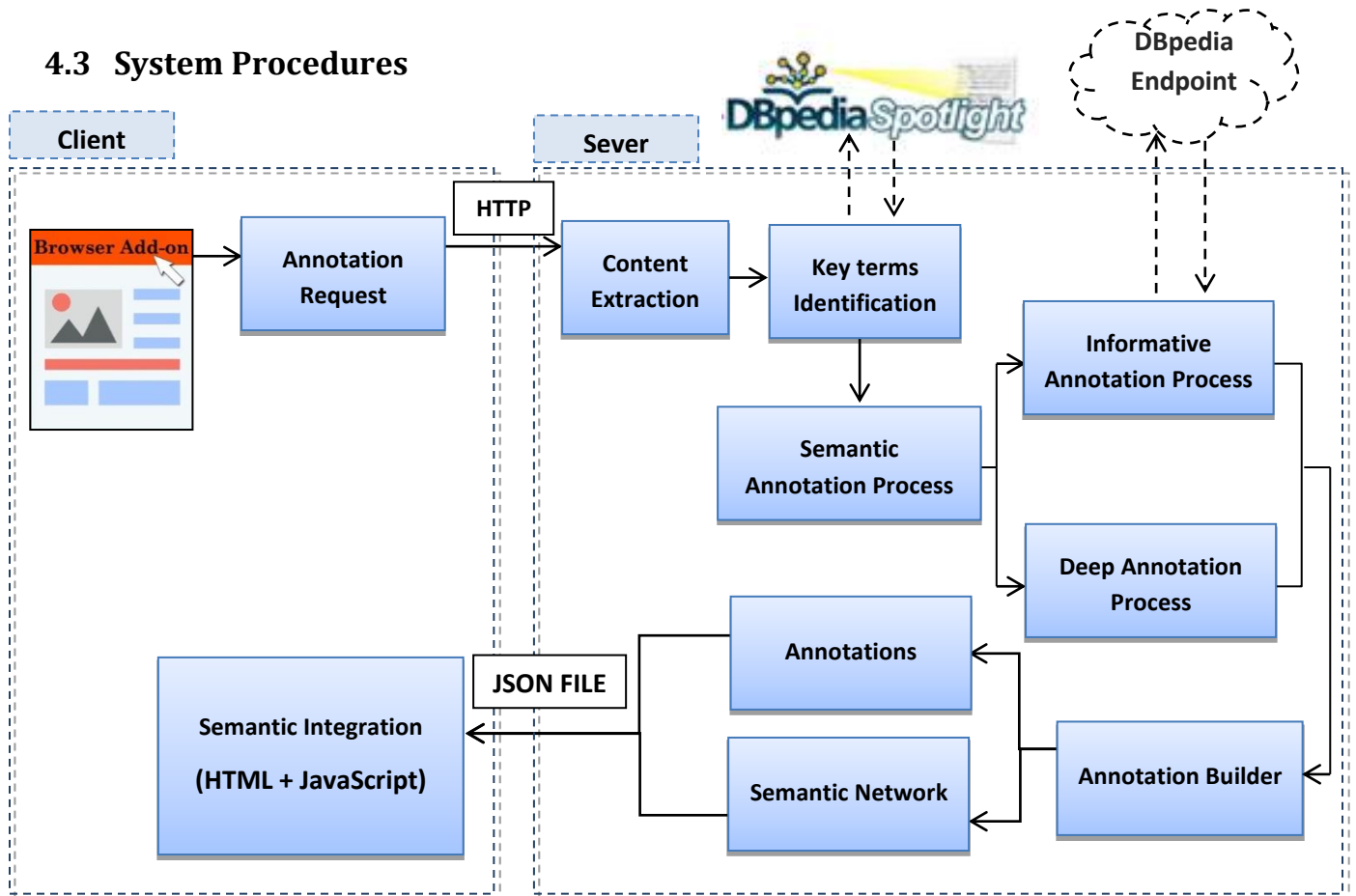


Figure 4-7: System Procedures

4.3.1 Server Side (RESTful Web Service)

On the server side, the RESTful web service receives a URL from the client, that represents the web page to be annotated.

As shown in Figure 4-7, the Server side consists of four modules: the Content Extraction module, the Entity Recognition Module, The Text Annotation Module and the Semantic Network Module. These module are explained in what follows:

4.3.1.1 Content Extraction

To determine the key terms in any web page, first of all, we need to extract the web page content as a plain text and prepare it to the next step, taking in consideration the need to remove all the unnecessary content such as footer, banner, advertisement... etc.

When the URL of the web page is received from the client, the Text Extraction module will extract the page content by using Boilerpipe API[44], which is Java library to extract the main content of the web page as a plain text and remove all unwanted html tags.

Any web page often display content in multiple parts such as the main area part (body), the navigation part, the banner part and the footer part. Our work focuses on the core area of the web page, the body, as the main content to be processed. We aim to extract the plain text and remove all non-necessary parts and elements such as HTML tags, navigation areas.. etc. as shown in Figure 4-8.


<p>Alstom: GE and French state agree to work on new offer</p> <p>© 20 June 2014 Business</p>  <p>Alstom makes the high-speed TGV train and is seen as one of France's key industrial firms</p> <p>The bid battle for France's Alstom has shifted in favour of General Electric.</p> <p>France's economy minister Arnaud Montebourg said that neither the GE bid, nor a joint Siemens-Mitsubishi Heavy Industries offer, met the government's demands.</p> <p>But the government had chosen to work on a new GE bid that would involve the French state buying a stake in Alstom.</p> <p>"The Siemens-MHI offer was serious but the government has made up its mind," Mr Montebourg told a news conference.</p>	<p>Alstom: GE and French state agree to work on new offer</p> <p>20 June 2014</p> <p>From the section Business</p> <p>Alstom makes the high-speed TGV train and is seen as one of France's key industrial firms</p> <p>The bid battle for France's Alstom has shifted in favour of General Electric.</p> <p>France's economy minister Arnaud Montebourg said that neither the GE bid, nor a joint Siemens-Mitsubishi Heavy Industries offer, met the government's demands.</p> <p>But the government had chosen to work on a new GE bid that would involve the French state buying a stake in Alstom.</p> <p>"The Siemens-MHI offer was serious but the government has made up its mind," Mr Montebourg told a news conference.</p>
--	---

Figure 4-8: : Extract Content from web page

4.3.1.2 Key Terms Identification

After the web page main content has been extracted as a plain text, the next step is to process this block of text to identify key terms that should map to DBpedia resources. These key terms will be annotated and associated with DBpedia URLs inside the Web page. In the following gives a brief overview on the DBpedia Spotlight.

A. DBpedia Spotlight and Named Entity Recognition

DBpedia Spotlight is a REST API based on DBpedia that detects mentions of DBpedia resources in text [45]. Its major aim is to recognize entities within a plain text and disambiguate their meaning, in order to establish the annotation process.

DBpedia spotlight works according to a set of sequential phases[46], see Figure 4-9:

1. **Phrase Recognition step:** process a given plain text to find significant substrings to be annotated (e.g. Name Entity Recognition).
2. **Candidate Selection:** map these substrings to all possible DBpedia resources
3. **Disambiguation:** choosing (ranking/classifying) the properly matched DBpedia resource for each substring according to its contextual meaning
4. **Tagging:** give the end user a chance to decide on the policy that best fits their needs.

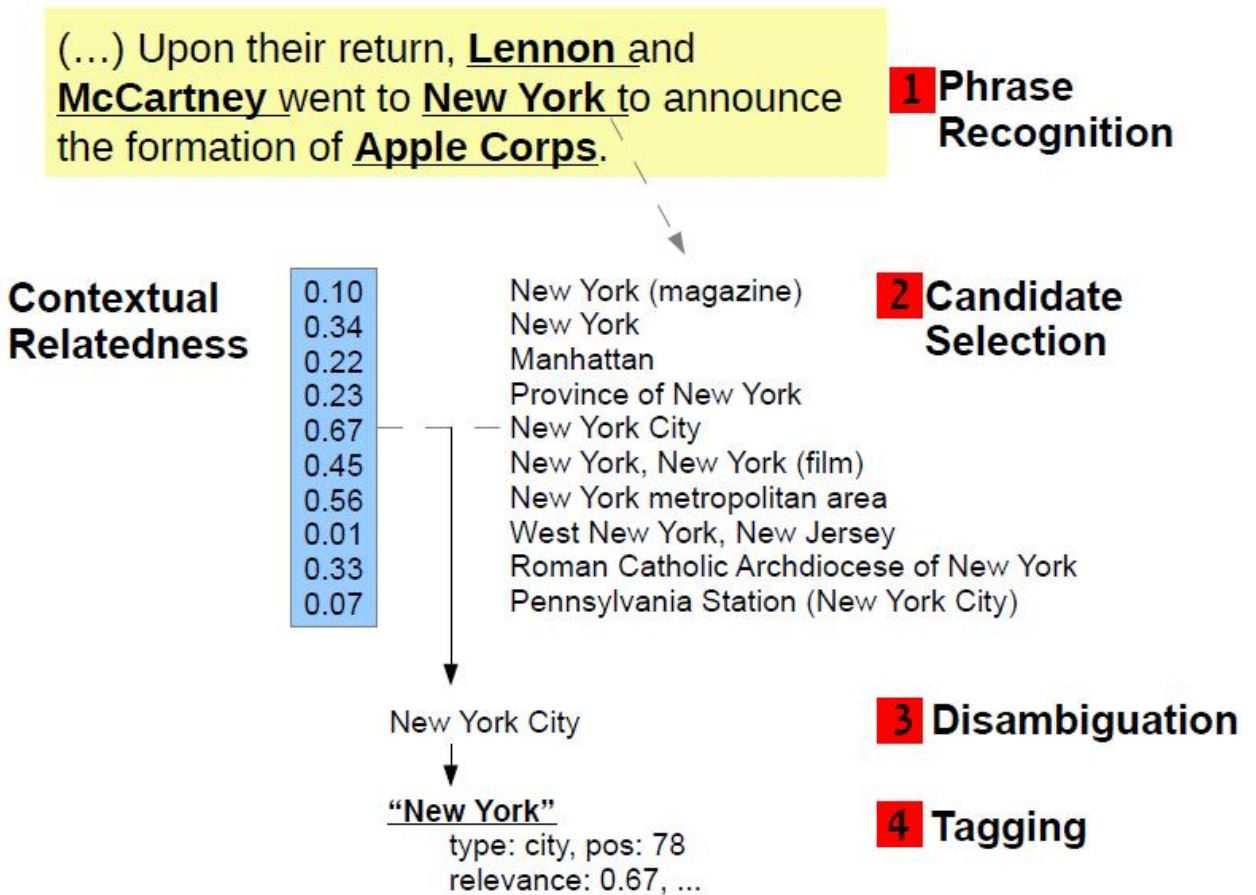


Figure 4-9: DBpedia Spotlight Default Workflow[46]

In our system, the extracted content of the page will be sent as a parameter to DBpedia Spotlight service[40], which automatically performs Named Entity Recognition (NER) and identifies terms that map to valid DBpedia URIs. NER is a subtask of information extraction that aimed to locate and classify atomic elements in text into predefined categories. It labels sequences of words in a text which are the names of things, such as person and company names, or gene[47].

B. DBpedia Spotlight Disambiguation Confidence

In Natural Language Processing (NLP), word disambiguation is the problem of determining which meaning of a word - that has more than one meaning - is suitable according to the occurrence of the word in a particular context . This process depends on many factors such as relevance of the word in a context and word contextual ambiguity; which means that there is two or more possible meanings within a single word.

DBpedia Spotlight system uses the context around the word, e.g. paragraphs, as information to find the most likely disambiguation, depending on the word occurrence on the text. Configureing DBpedia Spotlight is based on set of metrics (parameters) such as Disambiguation Confidence, which is ranging from 0 to 1[40]. If the confidence value is high ex. 0.7, 70% of incorrectly disambiguated words will be eliminated, which means only words with contextual ambiguity less than 0.3 (1-0.7) will be determined to be annotated.

Setting a high confidence value makes DBpedia Spotlight eliminate incorrect annotations as much as possible at the risk of losing some correct ones. Part of our evaluation will try to find the best confidence value to eliminate errors as much as possible, as will be mentioned in Section [6.4.1](#)

4.3.1.3 Semantic Annotation

In the previous stage, DBpedia spotlight returns only terms in the Web page that has DBpedia URIs. It does not return any information about the identified terms such as their definitions or relations to other terms. The following step is to perform semantic annotation by querying the DBpedia using SPARQL to extract detailed information about the terms identified by the DBpedia spotlight.

Semantic annotation is the process of enriching important terms in the text with semantic metadata extracted from DBpedia. It takes text as input and produces text with terms annotated as shown in Figure 4-10

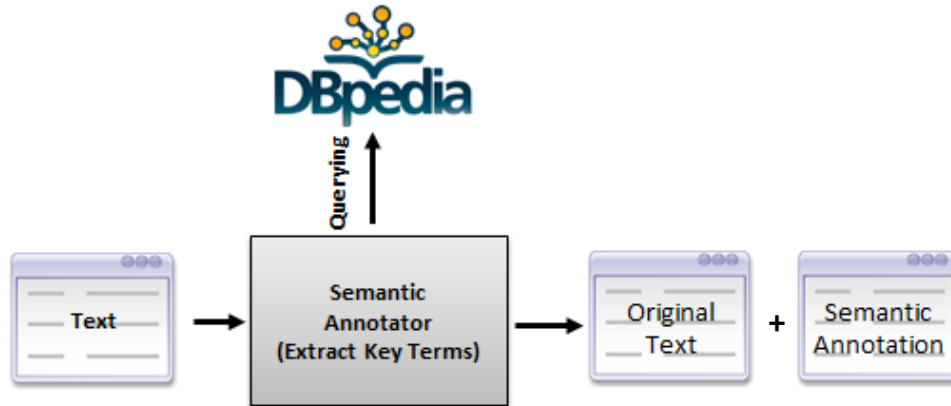


Figure 4-10: Semantic Annotation Process using DBpedia LOD

The semantic annotation consists of two major processes:

1. Informative annotation process: which aims to retrieve the major information, i.e. definition, related topics, etc., about the specific highlighted term that has been clicked.
2. Deep annotation process: which aims to find the relations – if exist - between different terms within the web page being browsed, and provide a set of related terms that do not appear in the web page content .

A. Informative Annotation Process

First step in semantic annotation is the Informative Annotation process which aims to extract a valuable and meaningful details about each DBpedia mentions resulted from the DBpedia spotlight annotation. This information shows a definition for each key term, a link to the Wikipedia article covering the term and other related terms. We used Apache JENA to build SPARQL query[15],to retrieve the major properties corresponding to these information from DBpedia dataset.

B. Deep Annotation Process

Deep annotation is formally defined as an annotation process that “utilizes information proper, information structures and information context in order to derive mappings between information structures”.[48]. This could be performed by finding the most possible relations between the terms (resources)within the web page, and other indirect terms to generate more knowledge. Then, these descriptive resources

are presented to the user in the form of a semantic network. The main objective of deep annotation process is to enrich user knowledge and experience, and provide extra information about web page key word, and about other related words that does not exist in the page.

Before explaining how deep annotation was performed, we provide an example showing the benefit of deep annotation to the end user. If we have a sport article with this sentence(see Figure4-11):

“ Messi travelled to London last night to receive his award”.

The key terms of this sentence are (Messi) and (London). These key terms can be annotated with the corresponding DBpedia entities [dbpedia:Lionel_Messi](http://dbpedia.org/page/Lionel_Messi) (http://dbpedia.org/page/Lionel_Messi) and [dbpedia:London](http://dbpedia.org/page/London) (<http://dbpedia.org/page/London>)

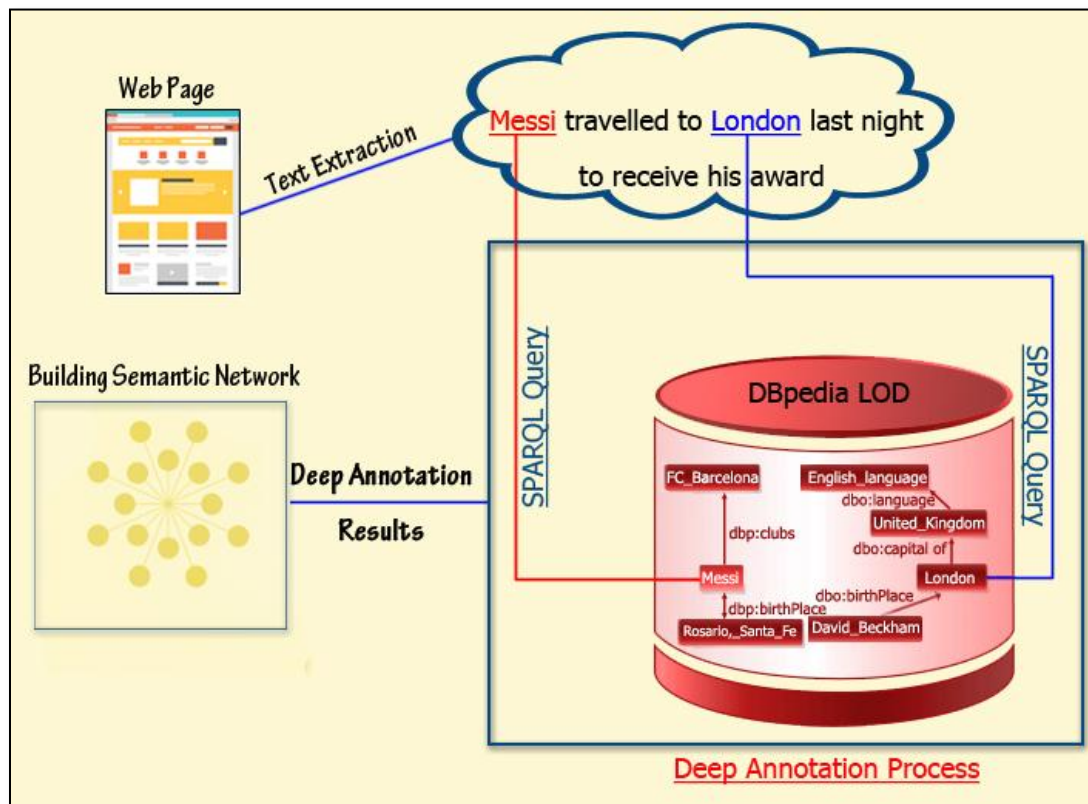
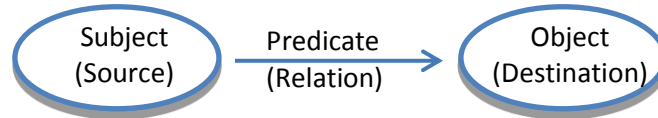


Figure 4-11: The Proposed Deep Annotation Approach

Using deep annotation process, some of additional DBpedia entities that are not directly observed in the sentence will be provided, such as [dbpedia:United_Kingdom](http://dbpedia.org/page/United_Kingdom) (http://dbpedia.org/page/United_Kingdom), which has a strong relation with London, and

[dbpedia:FC_Barcelona](http://dbpedia.org/page/FC_Barcelona) (http://dbpedia.org/page/FC_Barcelona), which has a strong relation with Messi, even if these two new terms (entities) are not in the same web page content. These DBpedia entities can be presented to user to act as complementary knowledge, allowing him/her to better understand the web page content. However, it is important to present these details in a user-friendly manner so that the user will not be cognitively overloaded. We aim to link and organize these entities using a visual graph that we called the “Semantic Network”.

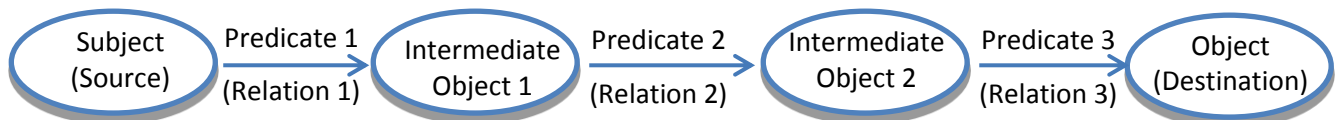
To find a relation between any two resources, first we have to determine the distance between them. Therefore, the relation between any DBpedia entities can be a direct relation as follow:



or indirect relations as follow:



When we try to find a direct relations between the words within the web page, sometimes we have a limit results and sometimes we have no results. This has motivated us to use the "deep annotation" process in order to find related topics from outside the page content. We used an incremental SPARQL query to find the relations between any key terms within the web page. The query returns results from both direct relations (a path with length one) and results from indirect relations (a path with maximum length of three) between the key terms.



Even more, the proposed deep annotation process find all possible relations between each one of the key terms and words that does not appear in the web page content, using an incremental SPARQL query that is similar to the previous query

As a result of the deep annotation process, many extra Dbpedia entities will be explored and added to the related set of terms. To limit that, we FILTER the results only with the most descriptive properties linked with DBpedia entities that are directly or indirectly associated with the basic key terms.

4.3.1.4 Annotation Builder

The results of both Informative annotation process and deep annotation process will be used to generate two major component, the core annotations of the web page content and the semantic network

A. Annotations

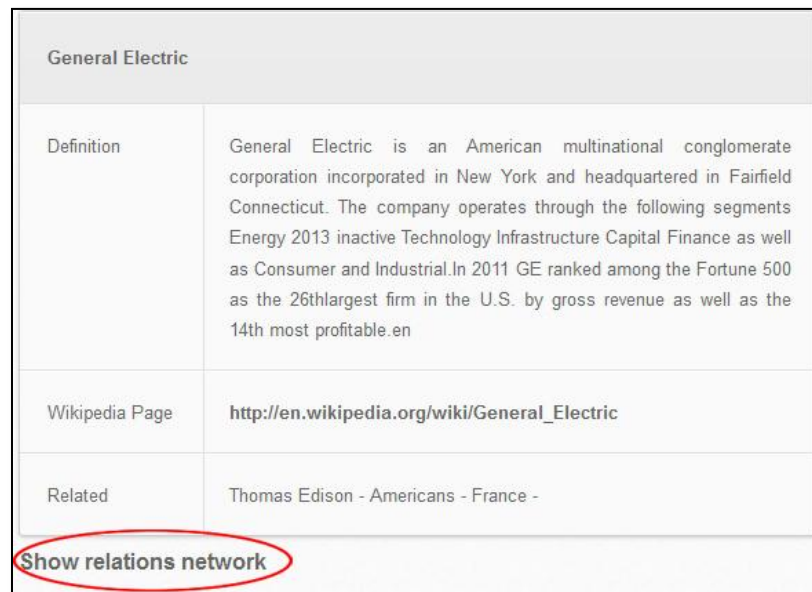
The annotations for each key term within the web page content are represented as a combination between the informative annotation process results and deep annotation process results. These annotations will be invoked from the client by using the AJAX function as will be discussed in Chapter 5.

B. Semantic Network

The major aim of semantic network is to illustrate visually the relations between the annotated DBpedia mentions within the web page in a way that a naïve user can easily understand and explore. The ultimate goal is to enhance the user experience and expand his/her knowledge. The semantic network displays and organizes different DBpedia entities that relate to each other.

To illustrate how the semantic network is constructed and used, assume that the Web page shown in Figure 4-2 (See Section 4.1) is annotated with DBpedia entities. As shown, the mentions of DBpedia entities are highlighted in a different color and converted to active links. For example, the word (France) in the page is annotated with metadata of the DBpedia entity [dbpedia:France](http://dbpedia.org/page/France)(<http://dbpedia.org/page/France>). Assume that the user wants to explore the relation between these highlighted terms, when the user clicks on the linked term in the page, the window shown in Figure 4-3

(See Section 4.1) is displayed. It shows the different properties related to this entity. For example, when click on the highlighted word (General Electric), a pop-up window is displayed contains General Electric's definition that has been extracted from DBpedia, Wikipedia article URL which is about General Electric (https://en.wikipedia.org/wiki/General_Electric), and a set of related words (Thomas Edison, Americans - France). The window also shows a link (see Figure 4-12), on which clicking will generated the graph in Figure 4-4.



General Electric	
Definition	General Electric is an American multinational conglomerate corporation incorporated in New York and headquartered in Fairfield Connecticut. The company operates through the following segments Energy 2013 inactive Technology Infrastructure Capital Finance as well as Consumer and Industrial. In 2011 GE ranked among the Fortune 500 as the 26th largest firm in the U.S. by gross revenue as well as the 14th most profitable.en
Wikipedia Page	http://en.wikipedia.org/wiki/General_Electric
Related	Thomas Edison - Americans - France -
Show relations network	

Figure 4-12: A hyperlinked text to generate the semantic network

The semantic network consists of two major parts: the network area and info-box area (See Figure 4-4, and Figure 4-5 Section 4.1). The network area shows the graphically illustration of the semantic network which includes a set of nodes and edges. The nodes represent the key terms from the web page and a set of terms resulted from the previous deep annotation process. The edges represent relation between these set words.

As shown in the graph, the graph is consist of a collection of orbits with a set of nodes located on it. Each node represents a DBpedia mention (key term).

To build the semantic network, two major steps will be performed, first determining the central node of the network, then building the semantic network structure.

a. Determine the Semantic Network Central Node

Suppose we have $[w_0, w_1, w_2, \dots, w_n]$ which represent a set of key words within the web page document that have been annotated with DBpedia URIs, and $[k_0, k_1, k_2, \dots, k_n]$ is the related words that do not appear in the web page content as mentioned in Section 4.2.3 and will be discuss in details in section 5.2.3:

1. First, we calculate $[w_{r0}, w_{r1}, w_{r2}, \dots, w_{rn}]$ which represent the number of relations assigned to each term with other terms depending on the previous semantic annotation step results. For example, if the word $[w_0]$ has a relations with $[w_1]$, $[w_2]$ and $[w_n]$, then $[w_{r0}]$ is 3 which is the number of relations associated to $[w_0]$.
2. Then, we select $[w_{center}]$ that has the largest number of relations to other words, and display it in the center of the network.
3. After determining the central node, other key words are set and organized on their orbits path around w_{center} according to the length of relation path between these words. Each orbit path O_i is located on it a set of words from inside and outside the web page content that is related to other words located on other orbits, where $O_i \in \{W_i\} \cup \{K_i\}$. Note that entities in the outer circles are less related to the central entity.

Algorithm 1 Finding the Central Node of the Network

1. **Input:** A list of words representing DBpedia mentions within the annotated web page $W[w_0, w_1, w_2, \dots, w_n]$, and $WR[w_{r0}, w_{r1}, w_{r2}, \dots, w_{rn}]$ which represent the number of relations assigned to each word.
2. **Output:** A central node $[w_{center}]$ which represent the word with the maximum number of relations.
3. $MAX = WR[0]$
4. $Wcenter = W[0]$
5. **for** (i=1 to n) **do**
6. {
7. **if** ($WR[i] > MAX$) **then**
8. {
9. $MAX = WR[i]$
10. $Wcenter = W[i]$

11. }
 12. **end if**
 13. }
 14. **end for**
 15. **return** Wcenter
-

b. Building the Semantic Network

After determining the central word, now we start to build the semantic network. It consist of two major elements:

1. Nodes (N) which is a set of key terms (represented by $[w_{r0}, w_{r1}, w_{r2}, \dots, w_{rn}]$) and its related words (represented by $[k_0, k_1, k_2, \dots, k_n]$).
2. Edges(E) which represent as the relations between different nodes.

Which makes the central node $W_{center} \in N$. this central node will be the root node and we will consider it as the root parent node.

Each node $\in N$ has a set of properties:

1. A unique id.
2. A data about this node, which contains its definition, its direct related terms as a children and the relations between them. Each one of these children is $\in N$, so each one of them considered as a new parent with its own properties such as the unique id, its children, and so on.

Algorithm 2 represent how to build the semantic network, with N representing the set of nodes (DBpedia mentions and its related terms), and E representing the edges (the relation between different nodes).

Algorithm 2 Building the Semantic Network

1. **Input:** Node list N set of key terms (represented by DBpedia mentions within the annotated web page $W[w_0, w_1, w_2, \dots, w_n]$) and its related words (represented by $K[k_0, k_1, k_2, \dots, k_n]$).
2. The central node $W_{center} \in N$
3. **Output:** A network represents relations between the web page key terms.
4. Create empty node list **Nodes** to store visited nodes

```
5.  while N is not empty do
6.    for each  $W_i \in N$  do
7.      if ( $W_i \notin Nodes$ ) then
8.        add ( $W_i, Nodes$ )
9.        if ( $W_i = W_{center}$ ) then
10.         CreateCenterNode ( $W_i$ )
11.        for each  $K_i \in K$  do
12.          CreateRelation( $W_i, K_i, E_i$ )
13.        end for
14.      end if
15.    end if
16.  end for
17. end while
```

4.3.2 Returning results to the Client Side

After annotations are extracted in the previous step, the next step aims to integrate the results retrieved from Server response with the content of web page being browsed. All the annotations and the semantic network will be represented as a JSON file that will returned to the client.

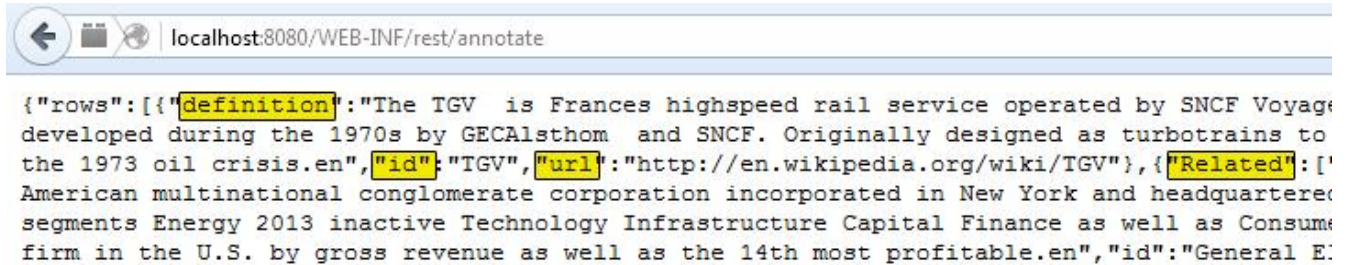
4.3.2.1 Semantic Integration

When the JSON file is received on the client, it will parsed to extract the annotations from it. The client side adds a new layout consisting of JavaScript functions and newly generated hyperlinks and injects them inside the Web page displayed on the browser. When click on any of these highlighted words, a new window opens containing information such as word definition, word's Wikipedia article URL, and a set of related words.

We chose to apply our system on Firefox web browser as a start. The plug-in appears as a button on the Firebox bar. If the button is clicked, a communication process will be established between the Client Side(the browser) and the Server side (RESTful web service). Through this connection, the client sends the URL of the web page

opened on the browser to the Server as a request. The Server performs the tasks, and sends the results as a response in JSON format. Then, the plug-in analyses the JSON file, extracts annotations and converts the keywords into active hyperlinks. By clicking these links, a new windows will appear containing the word definition, its Wikipedia URL , its related words, and a hyperlinked text to the illustrated semantic network (see Figure 4-13).

More details about how the integration process is established will be discussed in detail in Chapter 5



```
{ "rows": [ { "definition": "The TGV is Frances highspeed rail service operated by SNCF Voyage developed during the 1970s by GECAlsthom and SNCF. Originally designed as turbotrains to the 1973 oil crisis.en", "id": "TGV", "url": "http://en.wikipedia.org/wiki/TGV" }, { "Related": [ "American multinational conglomerate corporation incorporated in New York and headquartere segments Energy 2013 inactive Technology Infrastructure Capital Finance as well as Consum firm in the U.S. by gross revenue as well as the 14th most profitable.en", "id": "General E:
```

Figure 4-13: JSON representation results

4.4 Summary

In this chapter we discussed the proposed semantic annotation approach that aims to enhance the annotation process through the use of deep annotation. Unlike the usual annotation techniques, deep annotation aims to find more extended, correlated and indirectly observable entities even if these entities are not contained in the Web page.

The proposed approach consists of two main parts: the server side that is responsible for web page content extraction, entity recognition, semantic annotation and build the semantic network. The second part is the client side that is responsible for integrating the semantic annotation results with the original web page, using a Firefox plugin.

Chapter 5

Design and Implementation

5.1 Overview

This chapter presents the practical part of the thesis. We introduce the development processes and tools that have been used to accomplish the implementation of every step of our proposed approach. As mentioned in the previous chapter, the proposed system consists of two basic parts: Server Side and Client Side.

5.2 Server Side (RESTful Web Service)

The service side is responsible for handling the request for annotating the web page being browser by the user. It consists of several parts that process the content of the page, maps them with the DBpedia content, builds the annotations and links them to the content of the page. It returns to the user a JSON file containing all annotations. The server sides also builds the semantic graph that shows the semantic associations between the page's key terms as well as the deep annotations. The server side consists of the following module:

5.2.1 Content Extraction

When a request of annotation is sent by the client, its content will be parsed to get the body content of the web page, and returned as a plain text to be processed later in order to determine the key terms to be annotated. we used Boilerpipe, a Java library written by Christian Kohlschütter[44] to extract the main contents of the web page as a plain text and remove all unwanted html tags.

Step Result:	The main content of the web page as a string.
---------------------	---

5.2.2 Key Terms Identification

The next step is to preprocess the extracted text to determine the key terms in it. We used DBpedia Spotlight for this task. DBpedia Spotlight is an open tool in Java that is

designed to automatically detect mentions of DBpedia resources in a given text and annotate them with their corresponding DBpedia URIs[45].

The major benefit of DBpedia Spotlight is its ability to recognize entities within a plain text and disambiguate their meanings, in order to establish the annotation process depending on Named Entity Recognition technique.

(a) DBpedia Spotlight Disambiguation Confidence

As mentioned in Section 4.2.2, the confidence parameter ranging from 0 to 1. High confidence avoids incorrect annotations as much as possible, which means the number of extracted entities will be less than the number of extracted entities if the confidence value is low. For example, if we have the following sentence as an input to DBpedia spotlight:

“A finance ministry official said both genuine and forged passports were in the packets intercepted in the post”

if we set the confidence value to **0.3**, the extracted entities will be as follow (the weighted underlined words):

A **finance ministry official** said both **genuine** and **forged passports** were in the **packets intercepted** in the **post**.

and if we set the confidence value to 0.5, the extracted entities will be:

A finance ministry official said both genuine and forged **passports** were in the **packets** intercepted in the post.”.

The importance of identifying the appropriate value of the confidence lies in the error eliminated in the annotation process. In the previous example, and with confidence value = 0.3, the word genuine (*which means real or true*) has been selected as a key word, but when the annotation process is completed, its meaning refers to *Windows Genuine Advantage (WGA); an anti-piracy system created by Microsoft*.

As will be discussed later in [Chapter 6](#), we tested different values of confidence parameter to find the best value that can eliminate incorrect annotations as possible.

Using DBpedia Spotlight in this stage of our system, the result of this step is a set of words and their DBpedia URIs.

Step Result:

A set of DBpedia URIs that correspond to DBpedia mentions in the web page.

5.2.3 Semantic Annotation

The output of the previous phase is a set of DBpedia URIs that correspond to terms in the web page, e.g. the term France will be associated with the DBpedia URI: <http://dbpedia.org/page/France>. However, these URIs will not be understandable for naïve users. Therefore, we aim to extract extra details about these URIs. These details should be easily readable by users, and provide them with explanatory information about the term.

For each DBpedia URI retrieved from DBpedia spotlight, we extract the main properties by querying the DBpedia. These properties included the definition of the term, its link to Wikipedia article, and other related topics that will be used later to build the semantic graph. To perform this process, two major queries were built.

5.2.3.1 Informative Annotation Process

First step in semantic annotation is to extract a valuable information about each DBpedia resource and its properties by querying DBpedia to retrieve a meaningful information. Our SPARQL query (see Figure 5-1) retrieves three major properties for each resource: label, comment and isPrimaryTopicOf.

```
query = "PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> "  
+ "PREFIX foaf: <http://xmlns.com/foaf/0.1/>"  
+ "SELECT ?label ?abstract ?wikilink WHERE {"  
+ "<http://dbpedia.org/resource/" + list1.get(i) + "> rdfs:label ?label ."  
+ "<http://dbpedia.org/resource/" + list1.get(i) + ">" +  
+ "<http://www.w3.org/2000/01/rdf-schema#comment> ?comment." +  
+ "<http://dbpedia.org/resource/" + list1.get(i) + "> foaf:isPrimaryTopicOf ?wikilink." +  
+ "FILTER langMatches( lang(?comment), 'en')" +  
+ "FILTER langMatches( lang(?label), 'en')}";
```

Figure 5-1: Querying DBpedia to extract resource information

The property `rdfs:label` provides a human-readable information about a resource's name.

For example, for the DBpedia resource: http://dbpedia.org/resource/General_Electric the label will be General Electric.

The second property is `rdfs:comment`, and provides a human-readable description of a resource to clarify its meaning.

There is two DBpedia properties that can be used to provide a descriptive information for any resources, `dbo:abstract` and `rdfs:comment`. The key difference between them is that DBpedia abstracts usually include the first paragraph of a Wikipedia page as simple text. Comments are substrings of abstracts limited to two sentences as a short abstract. So it is more easier and more understandable for a naïve user to read a shorter description than longer one.

Figure 5-2 shows DBpedia `rdfs:comment` and `dbo:abstract` for DBpedia resource: <http://dbpedia.org/page/France>

<code>rdfs:comment</code>	<ul style="list-style-type: none">France (/fræns/; French: [fʁɑ̃s] (13px)), officially the French Republic (French: République française [ʁepyblik fʁɑ̃sɛz]), is a unitary sovereign state comprising territory in western Europe and several overseas regions and territories. Metropolitan France extends from the Mediterranean Sea to the English Channel and the North Sea, and from the Rhine to the Atlantic Ocean; France covers 640,679 square kilometres (247,368 sq mi) and has a population of 66.6 million.
<code>dbo:abstract</code>	<ul style="list-style-type: none">France (/fræns/; French: [fʁɑ̃s] (13px)), officially the French Republic (French: République française [ʁepyblik fʁɑ̃sɛz]), is a unitary sovereign state comprising territory in western Europe and several overseas regions and territories. Metropolitan France extends from the Mediterranean Sea to the English Channel and the North Sea, and from the Rhine to the Atlantic Ocean; France covers 640,679 square kilometres (247,368 sq mi) and has a population of 66.6 million. It is a semi-presidential republic with its capital in Paris, the nation's largest city and the main cultural and commercial center. The Constitution of France establishes the country as secular and democratic, with its sovereignty derived from the people. During the Iron Age, what is now France was inhabited by the Gauls, a Celtic people. The Gauls were conquered by the Roman Empire in 51 BC, which held Gaul until 486. The Gallo-Romans faced raids and migration from the Germanic Franks, who dominated the region for hundreds of years, eventually creating the medieval Kingdom of France. France has been a major power in Europe since the Late Middle Ages, with its victory in the Hundred Years' War (1337 to 1453) strengthening French state-building and paving the way for a future

Figure 5-2: DBpedia comment and a part of DBpedia abstract for a DBpedia resource

The last property is `foaf:isPrimaryTopicOf`, which is a property used to relate a DBpedia resource to its Wikipedia article. Providing a link to Wikipedia article allows the user to access more details about the pertinent topic.

These three properties provide a readable and meaningful information about each DBpedia mention in a way that the naïve user can understand easily. The previous DBpedia query restricts the results to English language only by using the FILTER clause.

```
"FILTER langMatches( lang(?comment), 'en')" +  
"FILTER langMatches( lang(?label), 'en')}";
```

5.2.3.2 Deep Annotation Process

After retrieving the required information for every resource, we try to find the relations between these resources which will be used for building the semantic network.

The relation between any two DBpedia resources may be determined directly. For example, and as shown on Figure 5-3, the relation between dbpedia:Alstom and dbpedia:France can be determined using the following query which consists of a single triple:

```
select * where  
{  
    <http://dbpedia.org/resources/Alstom> ?pre1 <http://dbpedia.org/resources/France>  
}
```

Figure 5-3: SPARQL query sample to retrieve relation between two resources with path length of one

Where ?pre1 is the predicate linking the two resources. However, sometimes queries like this may not return any result when the relation between resources is indirect and is achieved through a sequence of relations. So we built our SPARQL query to find indirect relations between any two resources with a path with length of one or more. The indirect relation can be found in multi-levels according to the path length between any DBpedia resources. E.g. if the path length between dbpedia:Alstom and dbpedia:France length of two, the query will be as shown in Figure 5-4:

```
select * where  
{  
    <http://dbpedia.org/resource/Alstom> ?pre1 ?obj1 .  
    ?obj1 ?pre2 <http://dbpedia.org/resource/France>.  
}
```

Figure 5-4: Simple SPARQL query to retrieve the relation between two resources with path length of two

And if the path length between them is length of two, the query will be as shown in Figure 5-5:

```

select * where
{
    <http://dbpedia.org/resource/Alstom> ?pre1 ?obj1 .
    ?obj1 ?pre2 ?obj2 .
    ?obj2 ?pre3 <http://dbpedia.org/resource/France>.
}
    
```

Figure 5-5: Simple SPARQL query to retrieve the relation between two DBpedia resources with path length of three

And so on. The longer the path between any two DBpedia mentions within the browsed web page, the less interested the user will be. So we choose to build a query that check all possible relations between any two DBpedia mentions with a path smaller than or equal to three. The incremental query resulted in a set of intermediate DBpedia resources between the source resource and the destination resource. These intermediate resources represent a part of our deep annotation process.

For example, When we executed the query shown in Figure 5-4, we have one intermediate DBpedia resource between dbpedia:Alstom and dbpedia:France, which will be the value of ?obj1, and the indirect relation between the two resources will be presented as ?pre1 and ?pre2. as shown in Figure 5-6

pre1	obj1	pre2
http://dbpedia.org/ontology/keyPerson	http://dbpedia.org/resource/Patrick_Kron	http://dbpedia.org/ontology/birthPlace

Figure 5-6: A sample relation between DBpedia resource Alstom and DBpedia resource France with path length of two

When we executed the query shown in Figure 5-5, we have two intermediate DBpedia resources between dbpedia:Alstom and dbpedia:France , which will be the value of ?obj1 and ?obj2. and the indirect relation between the two resources will be presented as ?pre1 , ?pre1 and ?pre3. as shown in Figure 5-7

pre1	obj1	pre2	obj2	pre3
http://dbpedia.org/ontology/keyPerson	http://dbpedia.org/resource/Patrick_Kron	http://dbpedia.org/ontology/birthPlace	http://dbpedia.org/resource/Paris	http://dbpedia.org/ontology/country

Figure 5-7: A sample relation between DBpedia resource Alstom and DBpedia resource France with path length of three

Furthermore, in case our SPARQL query did not return any results (the worst case) - which mean there is no direct or indirect relation between the two DBpedia mentions within the page- or if the returned results are not enough to enhance user browsing experience, we build another SPARQL query that find the related DBpedia resources to the DBpedia mentions within the web page. For Example, to find the related DBpedia resources to [dbpedia:France](http://dbpedia.org/resource/France) along a path of length of 2, we use the following query(see Figure 5-8):

```
SELECT ?pre1 ?obj1 ?pre2 WHERE {
<http://dbpedia.org/resource/France> ?pre1 ?obj1 .
?obj1 ?pre2 ?obj3
}
```

Figure 5-8: SPARQL query to retrieve the related resources to a given DBpedia resources with path length of two where ?obj2, is the related DBpedia resource to [dbpedia:France](http://dbpedia.org/resource/France), ?obj1 is the intermediate resource, and the indirect relation between the two resources will be presented as ?pre1 and ?pre2.

Sometimes, the previous query can return results that may disturb the user. For example, the previous query returned a non-English value for ?obj3 as shown in Figure 5-9

pre1	obj1	pre2	obj3
http://dbpedia.org/ontology/anthem	http://dbpedia.org/resource/La_Marseillaise	http://www.w3.org/2000/01/rdf-schema#label	"馬賽曲"@zh

Figure 5-9: A non-English results for a given Sparql Query

To resolve this problem, we used a set of filter expressions to restrict the results and remove the unwanted ones. For example, we restrict results to English language only, restrict the intermediate predicate to properties only, restrict the intermediate objects to resources only... etc.

Here is some of FILTER restrict conditions we use:

1. FILTER (langMatches(lang(?objlabel), 'en')

Returns true if language-tag (first argument) matches language-range (second argument). We use it to restrict the returned results to English language only.

2. FILTER (?pre1 != <http://www.w3.org/2002/07/owl#sameAs>)

owl:sameAs statement indicates that two URI references actually refer to the same thing: the individuals have the same "identity". We use it to eliminate duplicated properties results and duplicated DBpedia resources.

3. FILTER (?pre1 != <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>)

When we try to find the relation (predicate) between two entities (subject and object) we choose the property restrictions not to be `rdf:type`.

The above filters were carefully chosen to ensure that the query will generate appropriate results, and that the filters will not cause important results to be masked.

At the end, the result of Semantic Annotation process is represented as a JSON file containing each term, its related word, its definition, its label and its Wikipedia URL (see Figure 5-10).

```
crisis.en", "id": "TGV", "url": "http://en.wikipedia.org/wiki/TGV"}, {"Related": ["Thomas Edison", "Americans", "France"], "definition": "General Electric is an American multinational conglomerate corporation incorporated in New York and headquartered in Fairfield Connecticut. The company operates through the following segments Energy 2013 inactive Technology Infrastructure Capital Finance as well as Consumer and Industrial. In 2011 GE ranked among the Fortune 500 as the 26th largest firm in the U.S. by gross revenue as well as the 14th most profitable.en", "id": "General Electric", "url": "http://en.wikipedia.org/wiki/General_Electric"}, {"Related": ["Arrondissement of
```

Figure 5-10: JSON representation results

Figure 5-10 shows a snapshot of the JSON file showing the annotation of the term "General Electric". The JSON shows the term's extracted definition from DBpedia, its related Wikipedia article, and a set of related word. These related words include word from the same web page as well as other indirect related words (e.g. Thomas Edison, Americans) resulted from the deep annotation process that did not appear in the web page content.

Step Result:

A JSON file contains the key term definition, its Wikipedia URL and its related word if any.

5.2.4 Semantic Network

The main aim of the semantic network is to enhance user experience and knowledge by illustrating the relations between different words and their definitions. Not only terms from the web page will be shown in the network, but also other words that do not exist within the page but are related to its content.

The semantic network consist of two parts:

1. The network area, which shows a set of nodes and edges. Each node represents a word that could be one of the key words from the page or one of the related words not included in the page.

As described previously in Chapter 4, the central node of the network is the word with the maximum relations with other words. A three-level orbits are created around the node (see Figure 5-11). These three orbits correspond to the three levels of SPARQL queries in the deep annotation process. The first orbit contains nodes representing DBpedia entities that have direct relations with the central node (Path length = 1). The second orbit contains entities that have indirect relations through a path of length 2 with the central node. etc. as shown in Figure 5-1

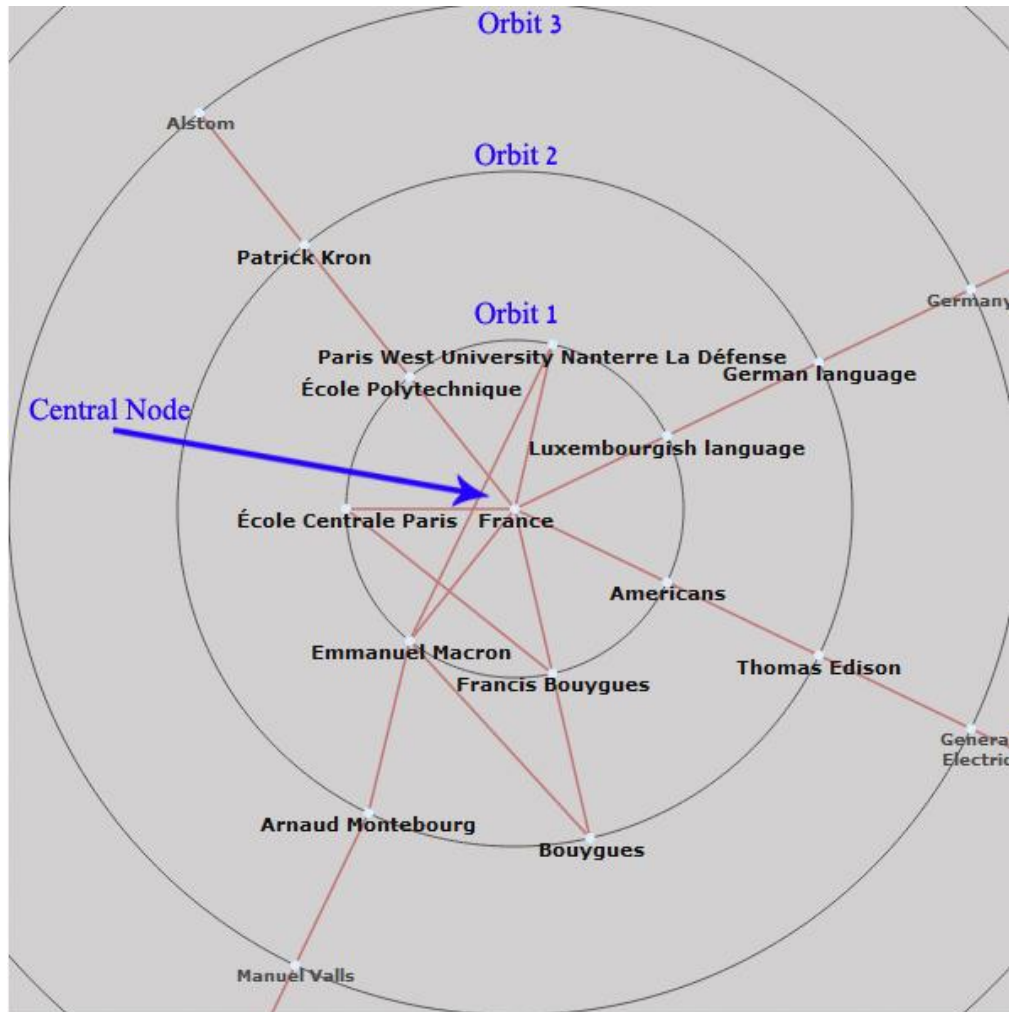


Figure 5-11: Central node and orbits around it

2. The info-box area, which contains information for each active node. For example, when clicking on the node titled "France", the info box shown in Figure 5-12 will be presented to the user. It shows more information about the term such as its related words with brief description on each.

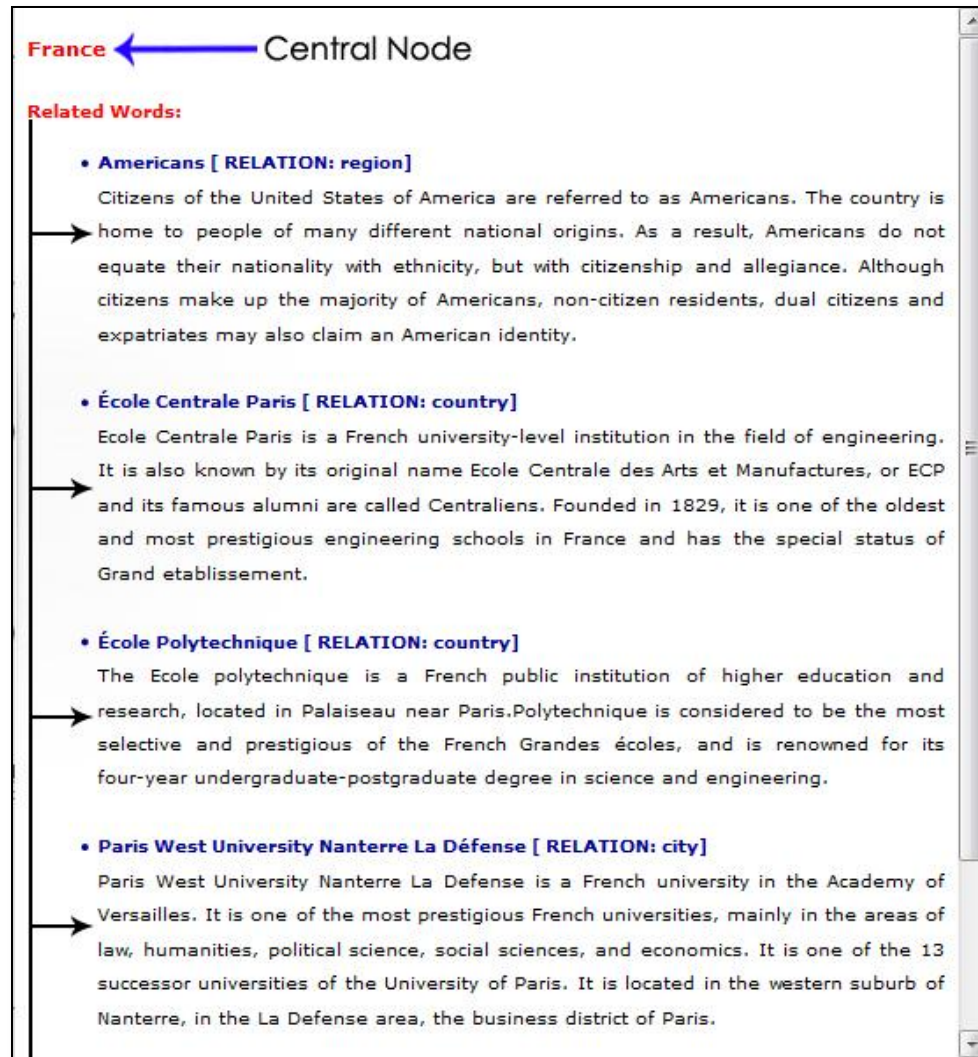


Figure 5-12: Info-box and related word information

When click on any node, the whole network will be restructured so that the new term converts to central node, and all other nodes are repositioned accordingly while preserving the relations between nodes.

To build this semantic network, we used JavaScript InfoVis Toolkit(JIT)[49]that provides tools for creating Interactive Data Visualizations on the Web.

5.3 Returning results to the Client Side

All information generating on the server will be returned to the client browser as a JSON file. On the client side, the JSON content will be converted to a layer of hyperlinks and JavaScript functions. This will cause the annotated terms to be highlighted in different

colors. When clicking on any of these highlighted words, a new window will open to show information about that word such as word definition, word's Wikipedia article URL, and a set of related words (see Figure 5-13). A new link will be also added to the web page to allow the user to view the semantic network.

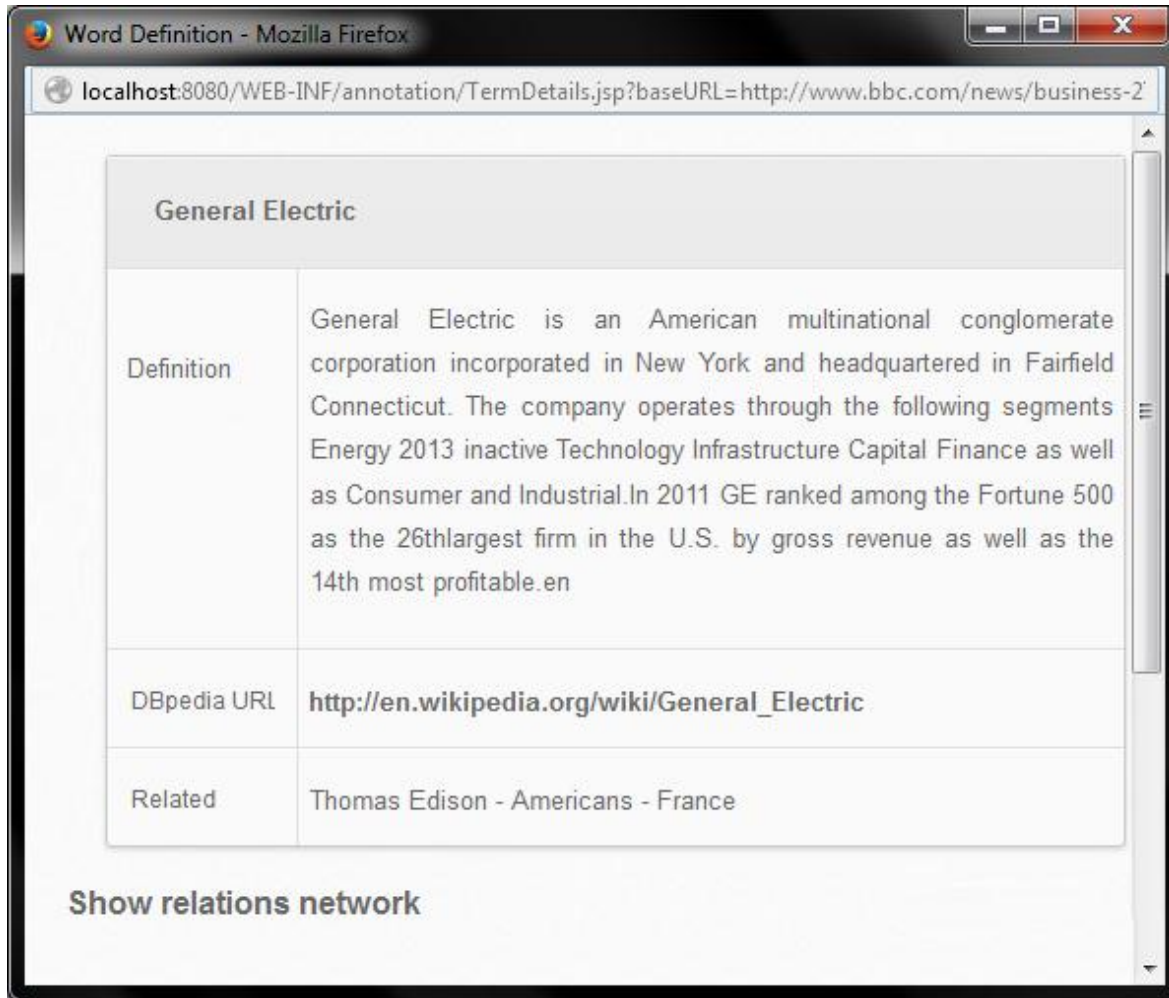


Figure 5-13: Annotated term details

5.3.1 Semantic Integration

The annotated service was integrated into a Firefox web browser. We built a Firefox add-on to achieve this purpose (see Figure 5-14).

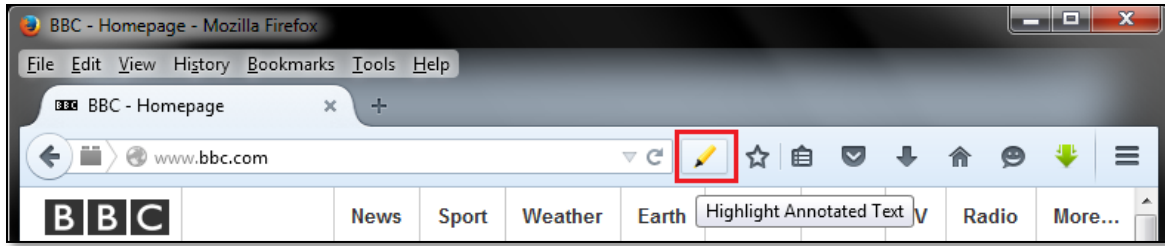


Figure 5-14: The proposed Firefox plugin icon

The add-on communicates with the server through a Restful web service. The web service is invoked from the client by using the AJAX function shown in Figure 5-15.

```
xmlhttp.open('GET', "http://localhost:8080/WEB-INF/rest/annotate?url="+window.content.  
document.URL, true);  
    xmlhttp.send(null);  
    xmlhttp.onreadystatechange = function() {  
        if (xmlhttp.readyState == 4 && xmlhttp.status == 200) {  
            var jsonResult = xmlhttp.responseText;  
            var parsedJSON = JSON.parse(jsonResult);
```

Figure 5-15: AJAX function used to invoke the Restful web service

As mentioned earlier, results from the server are retrieved in JSON format. Parsing the JSON file helps to access each element within it. The JSON file representation consist of four major elements:

1. id: which represent the key words label, ex. France.
2. definition: which represent the key word definition.
3. url: which represent the key word Wikipedia article URL.
4. Related: which represent the related words to each key word.

The id element is the key element in this process that is used in the find and replace mechanism. FindAndReplaceDOMText methods takes three major parameters:

```
findAndReplaceDOMText(re, window.content.document.body, parentA);
```

1. The parameter “re” is the regular expression to match. In this case, “re” represents a key words from the JSON file represented by “id” value, ex. TGV

```
localhost:8080/WEB-INF/rest/annotate
{"rows":[{"definition":"The TGV is Frances highspeed rail service operated by SNCF Voyages the longdis developed during the 1970s by GECAlsthom and SNCF. Originally designed as turbotrains to be powered by the 1973 oil crisis.en","id":"TGV","url":"http://en.wikipedia.org/wiki/TGV"}, {"Related":["Thomas Edison American multinational conglomerate corporation incorporated in New York and headquartered in Fairfield segments Energy 2013 inactive Technology Infrastructure Capital Finance as well as Consumer and Industr firm in the U.S. by gross revenue as well as the 14th most profitable.en","id":"General Electric","url"
```

2. The parameter “window.content.document.body” represents the content of web page being browsed.
3. The parameter “parentA” represents the replacement node within the browsed web page.

When the method finds the text “parentA” that matches the wanted word “value of id ex. TVG”, parentA is converted to a hyperlink that has a highlighted style.

```
var parentA = window.content.document.createElement('a');
parentA.href="javascript:void(0)";
parentA.setAttribute('id', 'parent_'+row.id);
parentA.className = 'text-highlighter-style';
parentA.style.backgroundColor =
    prefManager.getCharPref("extensions.texthighlighter.background");
```

When the algorithm finds the first match, it keeps looking for another matches, and then applies the replacement when a match is found, until the JSON file parsing is finished and highlight all matched key words (See Figure 5-16).



Figure 5-16: Annotated terms on a web page

5.4 Summary

In this chapter we discuss the technical implementation of our work. We divide it into two major parts with subtask to each one of these parts; server side (which includes text extraction, entity recognition, semantic annotation and semantic network) and client part which includes the semantic integration process. We illustrate the results of each subtask and how it used in other tasks.

Chapter 6

Evaluation

6.1 Overview

This chapter presents the testing and evaluation of our system. First we will illustrate the evaluation objectives and evaluation frameworks. Then we presents the evaluation process, results and discussion.

6.2 Evaluation Objective

The major objective of the evaluation process is to test the correctness of the semantic annotations added to the web page by our system. By correction we mean that the information associated with the retrieved annotation is valid and is related to the annotated term.

6.3 Evaluation Framework

6.3.1 Data Set

To evaluate our system, we choose three news articles from BBC News website. We decided to apply our system on a news website because we think that many internet users prefer to go online to search for any up-to-date news. The main content of our chosen three articles can be found in Appendix A.

6.3.2 Human Subjects

To test our work, we ask three different users to use our system and evaluate the results in different experiments. The three users are a Computer and Communication Engineer, a Computer Engineer and Information Technology Specialist. Therefore, all users are frequent Internet users.

6.4 Evaluation Process

The proposed system consist of two major parts, the first one is the semantic annotation process which aims to link between DBpedia mentions within the web page and there corresponding DBpedia extracted information. The second part is the use of deep annotation process to build a semantic network that illustrates the relations between different DBpedia resources that are related to the page content.

In the following we present the evaluation of the semantic annotation process and discuss its results. Afterwards, we present the evaluation of the deep annotation process and the semantic network.

6.4.1 Semantic Annotation Evaluation

As mentioned in Section 5.2.2, we use DBpedia Spotlight as a named entity recognition to match between DBpedia mentions within a given text, and their corresponding DBpedia URIs. DBpedia spotlight has a confidence parameter that ranges from 0 to 1, that is used as a disambiguation metric. Therefore, part of our evaluation aims to determine the best confidence parameter that gives the best annotation results with the least possible errors.

In this evaluation process, we tried four different values for the confidence parameter, and we run our system to evaluate the returned results in each case to determine which confidence value will return results with minimum annotation errors.

We tested the annotation process using the following values for the confidence parameter: {0.3, 0.5, 0.6, 0.7}

6.4.1.1 Results

We ask the three human subjects to assess the correctness of the annotation results for each confidence parameter. Table 6-1 shows a sample result for one web page that has been annotated with a specific confidence value. The results shown in the table include the annotated terms, the definition of each term as extracted from DBpedia. The column on the right shows the assessment of the user as "correct" or "incorrect". The full results from the three users with different confidence values can be found in Appendix.

Table 6-1: Extracted Key Term with their corresponding extracted DBpedia definition and the testing human subjects

Sentence		Germany seizes fake Syrian passports in asylum inquiry		
Key Word		Extracted DBpedia Definition		H. S
1.	Germany	Germany, officially the Federal Republic of Germany, is a federal parliamentary republic in western-central Europe. It consists of 16 constituent states, which retain limited sovereignty, and covers an area of 357,021 square kilometres with a largely temperate seasonal climate. Its capital and largest city is Berlin.		√
2.	Syrian	Syria, officially the Syrian Arab Republic, is a country in Western Asia, bordering Lebanon and the Mediterranean Sea to the west, Turkey to the north, Iraq to the east, Jordan to the south, and Israel to the southwest. Its capital Damascus is among the oldest continuously-inhabited cities in the world.		√
Sentence		German customs officers have seized packages containing Syrian passports and police suspect they are being sold illegally to asylum seekers		
Key Term		Extracted DBpedia Definition		H. S
3.	German	Germany, officially the Federal Republic of Germany, is a federal parliamentary republic in western-central Europe. It consists of 16 constituent states, which retain limited sovereignty, and covers an area of 357,021 square kilometres with a largely temperate seasonal climate. Its capital and largest city is Berlin.		√
4.	asylum seekers	A refugee is a person who is outside their home country because they have suffered persecution on account of race, religion, nationality, or political opinion; because they are a member of a persecuted social category of persons; or because they are fleeing a war. Such a person may be called an "asylum seeker" until recognized by the state where they make a claim. In 2014, Palestine, Syria, and Afghanistan were the largest source country of refugees.		√
Sentence		As refugees from the Syrian civil war, most have a right to asylum		
Key Term		Extracted DBpedia Definition		H. S
5.	Refugees	A refugee is a person who is outside their home country because they have suffered persecution on account of race, religion, nationality, or political opinion; because they are a member of a persecuted social category of persons; or because they are fleeing a war. Such a person may be called an "asylum seeker" until recognized by the state where they make a claim. In 2014, Palestine, Syria, and Afghanistan were the largest source country of refugees.		√
6.	Syrian civil war	The Syrian Civil War, also known as the Syrian Revolution, is an ongoing armed conflict taking place in Syria. The unrest began in the early spring of 2011 within the context of Arab Spring protests, with nationwide protests against President Bashar al-Assad's government, whose forces responded with violent crackdowns.		√
Sentence		The EU border agency Frontex says trafficking in fake Syrian passports has increased, notably in Turkey		
Key Term		Extracted DBpedia Definition		H. S

7.	EU	The European Union is a politico-economic union of 28 member states that are located primarily in Europe. The EU operates through a system of supranational institutions and intergovernmental negotiated decisions by the member states. The institutions are: the European Commission, the Council of the European Union, the European Council, the Court of Justice of the European Union, the European Central Bank, the Court of Auditors, and the European Parliament.	√
8.	Frontex	Frontex is the agency of the European Union that manages the cooperation between national border guards that is undertaken to secure the external borders of the union, including from illegal immigration, human trafficking and terrorist infiltration. The agency was established in 2004 and is headquartered in Warsaw, Poland.	√
Sentence		A Frontex official, Fabrice Leggeri, told French radio station Europe 1 that "people who use these fake passports mostly speak Arabic	
Key Term		Extracted DBpedia Definition	H. S
9.	French	French is a Romance language, belonging to the Indo-European family. It descended from the spoken Latin language of the Roman Empire, as did languages such as Italian, Portuguese, Spanish, Romanian, Catalan and others. Its closest relatives are the other langue's historically spoken in northern France and in southern Belgium, which French has largely supplanted.	√
10.	radio station	Radio broadcasting is a one-way wireless transmission over radio waves intended to reach a wide audience. Stations can be linked in radio networks to broadcast a common radio format, either in broadcast syndication or simulcast or both. Audio broadcasting also can be done via cable radio, local wire television networks, satellite radio, and internet radio via streaming media on the Internet.	√
11.	Europe 1	Europe 1, formerly known as Europe n° 1, is a privately owned radio network created in 1955. It is one of the leading French radio broadcasters and heard throughout France. The network is owned and operated by Lagardère Active, a subsidiary of the Lagardère Group.	√
12.	Arabic	Arabic is the Classical Arabic language of the 6th century and its modern descendants excluding Maltese. Arabic is spoken in a wide arc stretching across the Middle East, North Africa, and the Horn of Africa. Arabic belongs to the Afro-Asiatic family. The literary language, called Modern Standard Arabic or Literary Arabic, is the only official form of Arabic.	√
Sentence		They may come from North Africa, the Middle East, but they have the profile of economic migrants,"	
Key Term		Extracted DBpedia Definition	H. S
13.	North Africa	North Africa or Northern Africa is the northernmost region of Africa. Geopolitically, the United Nations definition of Northern Africa includes eight countries or territories; Algeria, Egypt, Libya, Mali, Morocco, Sudan, Tunisia. Algeria, Morocco, Tunisia, Libya and often Mauritania and Western Sahara are the Maghreb, while Egypt and Sudan comprise the Nile Valley.	√

14.	Middle East	The Middle East is a region centered on Western Asia and Egypt. The Eurocentric term is used as a synonym for Near East, in opposition to Far East. The corresponding adjective is Middle Eastern and the derived noun is Middle Easterner.	√
Sentence		Turkey is a major transit country for refugees	
Key Term		Extracted DBpedia Definition	H. S
15.	Turkey	Turkey, officially the Republic of Turkey, is a contiguous transcontinental parliamentary republic largely located in Western Asia with the portion of Eastern Thrace in Southeastern Europe.	√
Sentence		Syrian passports may be a shortcut to asylum for fraudulent claimants (AFP)	
Key Term		Extracted DBpedia Definition	H. S
16.	AFP	The Australian Federal Police is the federal police agency of the Commonwealth of Australia.	×
Sentence		A finance ministry official said both genuine and forged passports were in the packets intercepted in the post	
Key Term		Extracted DBpedia Definition	H. S
17.	Packets	A network packet is a formatted unit of data carried by a packet-switched network. Computer communications links that do not support packets, such as traditional point-to-point telecommunications links, simply transmit data as a bit stream. When data is formatted into packets, the bandwidth of the communication medium can be better shared among users than if the network were circuit switched. A packet consists of two kinds of data: control information and user data (also known as payload).	×

For each experiment, we calculate the Precision value to determine the accuracy of extracted key terms from each web page as the following:

$$\begin{aligned}
 & \textit{Precision (P)} \\
 & = \frac{|\{\textit{relevant of annotations}\} \cap \{\textit{retrieved annotations}\}|}{|\{\textit{retrieved annotations}\}|} \qquad \text{Eq.(6.1)}
 \end{aligned}$$

Where ***relevant of annotations*** represents key terms that has been annotated correctly, and ***retrieved annotations*** represents the total number of the annotated key terms, whether they have been annotated correctly or not.

While we have no golden standard to determine the major key terms in any given text, we think that we cannot calculate the recall value in each experiment.

Table 6-2 shows the average Precision in the experiments.

Table 6-2: The average Accuracy

	The average Accuracy
Confidence = 0.3	51.83%
Confidence = 0.5	90.59%
Confidence = 0.6	88.28%
Confidence = 0.7	89.74%

6.4.1.2 Discussion

As shown in Table 6-1, 17 words were annotated from the web page, and their corresponding extracted DBpedia definitions are shown. Two of these annotations were marked as "incorrect" by the human subject. These terms are:

Sentence	Syrian passports may be a shortcut to asylum for fraudulent claimants (<u>AFP</u>)	
Key Term	Extracted DBpedia Definition	H. S
AFP	The Australian Federal Police is the federal police agency of the Commonwealth of Australia.	×

Our judge said that this is the wrong match between the key term and the definition. He said that AFP stands for the Agency of France Press not the Australian Federal Police.

Sentence	A finance ministry official said both genuine and forged passports were in the <u>packets</u> intercepted in the post	
Key Term	Extracted DBpedia Definition	H. S
Packets	A network packet is a formatted unit of data carried by a packet-switched network. Computer communications links that do not support packets, such as traditional point-to-point telecommunications links, simply transmit data as a bit stream. When data is formatted into packets, the bandwidth of the communication medium can be better shared among users than if the network were circuit switched. A packet consists of two kinds of data: control information and user data.	×

This is another wrong match between the key term and the definition. Packets in this sentence refers to a ship traveling at regular intervals between two ports.

By reviewing our testing users judgments, we found that the resulted wrong annotations between the key term and its definition depend on the word disambiguate and the misunderstanding of the contextual meaning of the key term.

Referring to Table 6-2, we found that the best annotation results were obtained when the confidence parameter was 0.5. We did further testing to tune the confidence parameter, and found that the best results can be obtained when the confidence parameter is equal to 0.53 with average accuracy of 94.12%. Therefore, we used this value in the subsequent experiments.

6.4.2 Deep Annotation and Semantic Network

As mentioned in Section 5.2.3 and Section 5.2.4, the deep annotation process aims to find the most possible relations between the DBpedia mentions within the web page content, and other indirect terms to generate more knowledge. To evaluate the semantic network, we constructed a semantic network for a single web page and asked the three subjects to assess the correctness of the terms included in the network. A snapshot of the assessed network is shown in Figure 4-4. The network consists of 23 different key terms and 18 relations in between.

6.4.2.1 Results

After asking our judges to read carefully the provided information about key terms and the relations between them, and justify the accuracy of these information by referring to Wikipedia web page for each presented key term, we found that. On average, 95.44% of the retrieved key terms and relations were assessed as correct.

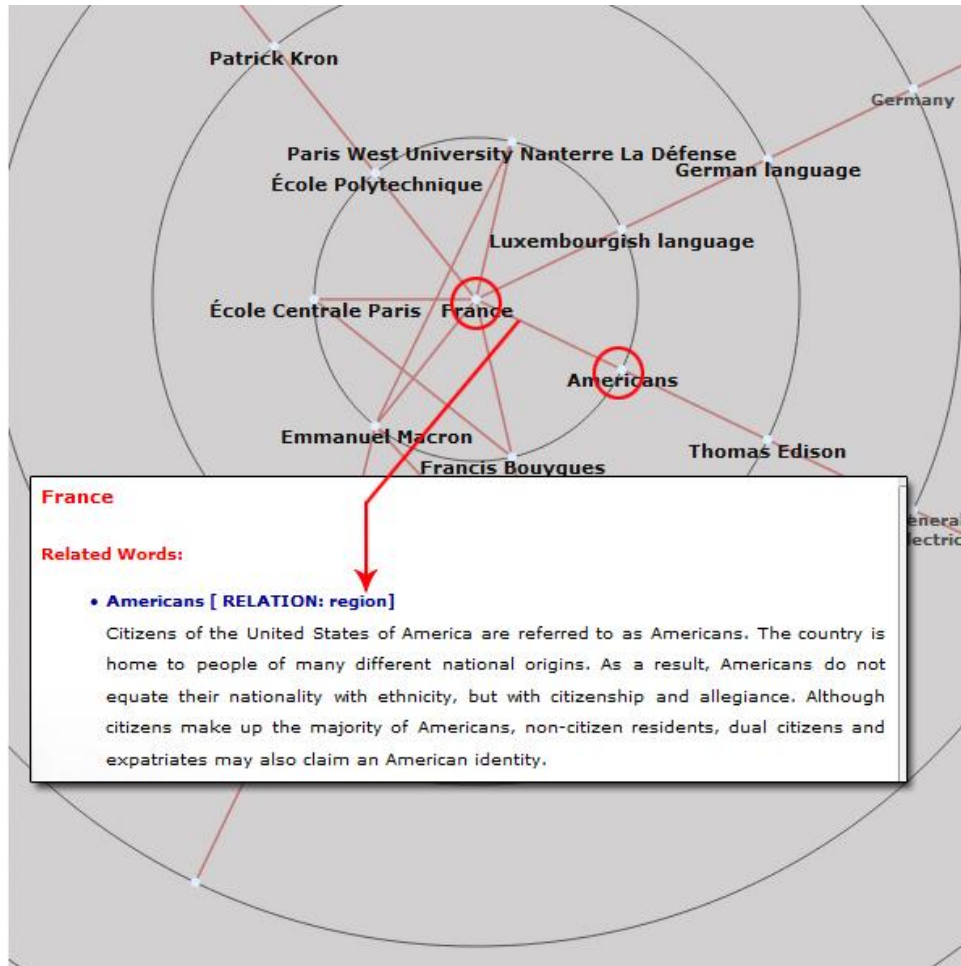


Figure 6-1: Relation caption between France and Americans

6.4.2.2 Discussion

In the following we discuss some of the incorrect results retrieved by the system and identify their causes. Some of the relations were incorrect. For example, the retrieved relation between (France) and (Americans) is (region) (see Figure 6-1). Our judges said that the region of the Americans is the United State of America not France. To determine the cause of this error, we referred to the France Wikipedia article, and we found no information in its info-box about Americans, but within the article, the word "Americans" is mentioned two times in two different places: the first place is for describing the relation between French spiritualist thinkers and some of the Americans ones.

In the early 20th century, French spiritualist thinkers such as Maine de Biran, Henri Bergson and Louis Lavelle influenced Anglo-Saxon thought, including the Americans Charles Sanders Peirce and William James, and the Englishman Alfred North Whitehead. In the late 20th century, partly influenced by German

The second time was when the article mentioned Louis XVI, support to the Americans.

Louis XVI, Louis XV's grandson, actively supported the Americans, who were seeking their independence from Great Britain (realized in the Treaty of Paris (1783)). The

6.5 Summary

In this chapter we discussed the testing and evaluation process to our system. We have three data sets that have been processed and annotated in four different experiments. In each experiment we set a different value to DBpedia spotlight confidence parameter, and compare the results accuracy with our work results accuracy.

We also discuss the accuracy of the information that has been provided by deep annotation process and the semantic network.

Chapter 7

Conclusion and Future Work

7.1. Conclusion

In this research, we introduced an approach to enhance web browsing experience for any naïve users using deep annotation and semantic visualization. This semantic annotation process will provide to the user all information he/she needs about the content of the web page, without the need to search over the internet.

The developed system processes any web page content, illustrates its key terms, and creates a link between them and their semantic metadata extracted from DBpedia LOD. It seeks additional techniques to the traditional semantic annotation process to make the annotation more constructive for Web browsing by using deep annotation process, which aims to find more extended, correlated and indirectly observable entities even if these entities are not contained in the Web page.

The developed system also provides a semantic network that visualizes the relationships between the different terms (entities) included in the Web page being browsed, in a way that could help the user better interpret the Web page content and utilize semantic annotations to gain broader knowledge. Our proposed annotation process was assessed by three human subjects, and results showed that 94.12% of the retrieved annotations were correct. Results also indicated that 95.44% of the terms included in the constructed semantic network was correct.

7.2. Future Work

As a future work, we will work to enhance the system reliability and improve the accuracy of our results. We will try to extend the annotation process and retrieve more details by exploiting Wikipedia in addition to DBpedia because Wikipedia has a wider coverage. After enhancing the annotation processing we will try to make our system compatible with any type of internet browser and not only Firefox browser.

Further study of this issue would be of interest of how to enhance user browsing experience in web sites with Arabic content. Semantic Annotation processes in Arabic language relays to limited re-built ontologies, so it will be a necessary to enrich Arabic language in DBpedia as a structured representation of Wikipedia in order to evaluate our system on web pages with Arabic content.

References

1. Uren, V. and M. Keynes, *Building Semantic Intranets: What is Needed in the Annotation Toolbox?*
2. Berners-Lee, T., J. Hendler, and O. Lassila, *The semantic web*. Scientific american, 2001. **284**(5): p. 28-37.
3. Oren, E., et al., *What are semantic annotations*. Relatório técnico. DERI Galway, 2006.
4. Quan, D.A. and R. Karger, *How to make a semantic web browser*, in *Proceedings of the 13th international conference on World Wide Web*. 2004, ACM: New York, NY, USA. p. 255-265.
5. Bertini, M., et al., *Web-based semantic browsing of video collections using multimedia ontologies*, in *Proceedings of the international conference on Multimedia*. 2010, ACM: Firenze, Italy. p. 1629-1632.
6. Grassi, M., et al., *Pundit: augmenting web contents with semantics*. Literary and linguistic computing, 2013: p. fqt060.
7. Mirizzi, R., et al., *Semantic wonder cloud: exploratory search in DBpedia*. 2010: Springer.
8. Chapman, S., B. Norton, and F. Ciravegna. *Armadillo: Integrating knowledge for the semantic web*. in *Proceedings of the Dagstuhl Seminar in Machine Learning for the Semantic Web*. 2005.
9. Garcia, J., U. Gut, and A. Galves. *Vocale-a semi-automatic annotation tool for prosodic research*. in *Speech Prosody 2002, International Conference*. 2002.
10. Dzbor, M., J. Domingue, and E. Motta, *Magpie-towards a semantic web browser*, in *The Semantic Web-ISWC 2003*. 2003, Springer. p. 690-705.
11. Sah, M., W. Hall, and D. De Roure, *SemWeB Semantic Web Browser-Improving Browsing Experience with Semantic and Personalized Information and Hyperlinks*. 2009.
12. Huynh, D., S. Mazzocchi, and D. Karger, *Piggy bank: experience the semantic web inside your web browser*, in *Proceedings of the 4th international conference on The Semantic Web*. 2005, Springer-Verlag: Galway, Ireland. p. 413-430.
13. *W3C Semantic Web*. [cited 2015 10 May]; Available from: <http://www.w3.org/standards/semanticweb/>.
14. Berners-Lee, T. *Artificial Intelligence and the Semantic Web*. 18 July 2006 [cited 2015 4 October]; Available from: <http://www.w3.org/2006/Talks/0718-aaai-tbl/>.

15. SPARQL Query Language for RDF. 15 January 2008 [cited 2015 10 September]; Available from: <http://www.w3.org/TR/rdf-sparql-query/>.
16. Paolo, C., O. Marco, and C. Tim, *DOMEO: a web-based tool for semantic annotation of online documents*.
17. Valkeapää, O. and E. Hyvönen. *A browser-based tool for collaborative distributed annotation for the semantic web.*(September 26 2006) *5th International Semantic Web Conference*. in *Semantic Authoring and Annotation Workshop*. 2006.
18. Toussaint, Y. *Semantic Annotation of Texts*. 2006. Invited talk at the 2nd Intern'l workshop on Enterprises and Networked Enterprises Interoperability, Vienna.
19. Ahmed, Z., T. Dandekar, and S. Majeed, *Review: Semantic Web; Ontology Specific Languages for Web Application Development*. IJWA, 2012. **4**(1): p. 33-41.
20. Miller, E., *An introduction to the resource description framework*. Bulletin of the American Society for Information Science and Technology, 1998. **25**(1): p. 15-19.
21. Arenas, M., C. Gutierrez, and J. Pérez, *On the Semantics of SPARQL*, in *Semantic Web Information Management*, R. de Virgilio, F. Giunchiglia, and L. Tanca, Editors. 2010, Springer Berlin Heidelberg. p. 281-307.
22. XML Soap. [cited 2015 4 October]; Available from: http://www.w3schools.com/xml/xml_soap.asp.
23. SOAP. 23 September 2015 [cited 2015 4 October]; Available from: <https://en.wikipedia.org/wiki/SOAP>.
24. Fielding, R.T., *Architectural styles and the design of network-based software architectures*. 2000, University of California, Irvine.
25. Nagarajan, M., *Semantic annotations in web services*, in *Semantic Web Services, Processes and Applications*. 2006, Springer. p. 35-61.
26. Slimani, T., *Semantic annotation: The mainstay of semantic web*. arXiv preprint arXiv:1312.4794, 2013.
27. Ley, M. *The DBLP computer science bibliography: Evolution, research issues, perspectives*. in *String Processing and Information Retrieval*. 2002. Springer.
28. Giles, C.L., K.D. Bollacker, and S. Lawrence. *CiteSeer: An automatic citation indexing system*. in *Proceedings of the third ACM conference on Digital libraries*. 1998. ACM.
29. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*. Nucleic acids research, 2003. **31**(1): p. 365-370.
30. Berners-Lee, T., C. Bizer, and T. Heath, *Linked data-the story so far*. International Journal on Semantic Web and Information Systems, 2009. **5**(3): p. 1-22.

31. *DBpedia* 2015 [cited 2014 10 December]; Available from: <http://dbpedia.org/About>.
32. Lehmann, J., et al., *DBpedia-a large-scale, multilingual knowledge base extracted from wikipedia*. *Semantic Web Journal*, 2014. **5**: p. 1-29.
33. *Virtuoso SPARQL Query Editor*. 2015 [cited 2015 6 Septmber]; Available from: <http://dbpedia.org/sparql>.
34. Yoo, S., Y. Kim, and S. Park, *An Educational Tool for Browsing the Semantic Web*. *Informatics in Education-An International Journal*, 2013(Vol12_1): p. 143-151.
35. Schuhmacher, M. and S.P. Ponzetto. *Exploiting dbpedia for web search results clustering*. in *Proceedings of the 2013 workshop on Automated knowledge base construction*. 2013. ACM.
36. Lama, M., et al., *Semantic linking of learning object repositories to DBpedia*. *Journal of Educational Technology & Society*, 2012. **15**(4): p. 47-61.
37. Lukovnikov, D., M. Verbeke, and B. Berendt. *User interest prediction for tweets using semantic enrichment with DBpedia*. in *BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial Intelligence, Delft, The Netherlands, November 7-8, 2013*. 2013. Delft University of Technology (TU Delft); under the auspices of the Benelux Association for Artificial Intelligence (BNVKI) and the Dutch Research School for Information and Knowledge Systems (SIKS).
38. Kobilarov, G., et al., *Media meets semantic web-how the BBC uses DBpedia and linked data to make connections*, in *The semantic web: research and applications*. 2009, Springer. p. 723-737.
39. Becker, C. and C. Bizer, *DBpedia Mobile: A Location-Enabled Linked Data Browser*. LDOW, 2008. **369**.
40. Mendes, P.N., et al. *DBpedia spotlight: shedding light on the web of documents*. in *Proceedings of the 7th International Conference on Semantic Systems*. 2011. ACM.
41. Zhang, Y., G. Cheng, and Y. Qu. *RelClus: Clustering-based Relationship Search*. in *International Semantic Web Conference (Posters & Demos)*. 2013.
42. Heim, P., et al., *RelFinder: Revealing relationships in RDF knowledge bases*, in *Semantic Multimedia*. 2009, Springer. p. 182-187.
43. *Apache Jena, A free and open source Java framework for building Semantic Web and Linked Data applications*. 2015 [cited 2014 8 March]; Available from: <http://jena.apache.org>

44. Kohlschütter, C., P. Fankhauser, and W. Nejdl. *Boilerplate detection using shallow text features*. in *Proceedings of the third ACM international conference on Web search and data mining*. 2010. ACM.
45. Daiber, J., et al. *Improving efficiency and accuracy in multilingual entity extraction*. in *Proceedings of the 9th International Conference on Semantic Systems*. 2013. ACM.
46. Mendes, P.N., *Adaptive Semantic Annotation of Entity and Concept Mentions in Text*. 2013, Wright State University.
47. *Natural language processing*. 5 September 2015 [cited 2015 5 March]; Available from: https://en.wikipedia.org/wiki/Natural_language_processing#Major_tasks_in_NLP.
48. Handschuh, S., S. Staab, and R. Volz. *On deep annotation*. in *Proceedings of the 12th international conference on World Wide Web*. 2003. ACM.
49. Belmonte, N.G. *JavaScript InfoVis Toolkit*. 2013 2013 [cited 2015 8 September]; Available from: <http://philogb.github.io/jit/>.

A. Appendix A

A.1 Experimental Testing

A.1.1 Text Extraction

Table A-1: Extracted Data Set

First Data Set: http://www.bbc.co.uk/news/world-europe-34150408
<p>Germany seizes fake Syrian passports in asylum inquiry 4 September 2015 Image copyright AFP Image caption Syrian passports may be a shortcut to asylum for fraudulent claimants German customs officers have seized packages containing Syrian passports and police suspect they are being sold illegally to asylum seekers. A finance ministry official said both genuine and forged passports were in the packets intercepted in the post. Germany is letting Syrians register for asylum regardless of where they entered the EU. As refugees from the Syrian civil war, most have a right to asylum. The passports can help fraudulent claimants to get asylum, the EU says. The ministry official declined to say how many Syrian passports had been found in the customs checks. The German police are now investigating. The EU border agency Frontex says trafficking in fake Syrian passports has increased, notably in Turkey. A Frontex official, Fabrice Leggeri, told French radio station Europe 1 that "people who use these fake passports mostly speak Arabic. "They may come from North Africa, the Middle East, but they have the profile of economic migrants," he said. Germany has by far the highest number of asylum applicants in the EU, many of them Syrians and Afghans, but many also from the western Balkan countries. Turkey is a major transit country for refugees and other migrants heading for the EU, and is also housing more than two million Syrian refugees in camps.</p>
Second Data Set: http://www.bbc.com/news/world-middle-east-34576035
<p>Hajj deaths 'almost triple' official Saudi toll 19 October 2015 From the section Middle East Image copyright AP</p>

Table A-1: Extracted Data Set

Image caption Pilgrims were crushed to death in Mina when two large crowds met
A crush near Mecca last month killed nearly three times as many people as Saudi Arabia has admitted, according to a tally by the Associated Press (AP).
AP said on Monday that at least 2,110 people died in the tragedy at the annual Hajj pilgrimage - far more than the official Saudi death toll of 769.
The new figure comes from media reports and statements from 30 countries who lost citizens, AP said.
The crush was the deadliest incident to strike the Hajj in 25 years.
Saudi officials have not updated their death toll - or the number of injured, which stands at 934 - since 25 September.
Iran says it lost 465 of its citizens, making it the worst affected nation. Many of the dead also came from Africa: Nigeria said it lost 199 people, while Mali lost 198, and Egypt 192, according to the AP count.
The AP tally comes after Saudi officials said they held a meeting about the disaster late on Sunday night.
According to the country's state press agency, SPA, Crown Prince Mohammed bin Naif bin Abdul Aziz, who is also the kingdom's interior minister, oversaw the meeting.
An investigation into the incident ordered by King Salman is ongoing.
"The crown prince was reassured on the progress of the investigations," the SPA report said.
Previously, the deadliest incident at the Hajj was a 1990 stampede that killed 1,426 people.

Third Data Set: <http://www.bbc.co.uk/news/health-34615621>

Processed meats do cause cancer – WHO

26 October 2015

Image copyright Thinkstock

Processed meats - such as bacon, sausages and ham - do cause cancer, according to the World Health Organization (WHO).

Its report said 50g of processed meat a day - less than two slices of bacon - increased the chance of developing colorectal cancer by 18%.

Meanwhile, it said red meats were probably carcinogenic but there was limited evidence.

The WHO did stress that meat also had health benefits.

Cancer Research UK said this was a reason to cut down rather than give up red and processed meats.

And added that an occasional bacon sandwich would do little harm.

What is processed meat?

Processed meat has been modified to either extend its shelf life or change the taste and the main methods are smoking, curing, or adding salt or preservatives.

Simply putting beef through a mincer does not mean the resulting mince is processed unless it is modified further.

Processed meat includes bacon, sausages, hot dogs, salami, corned beef, beef jerky and ham as well as canned meat and meat-based sauces.

Table A-1: Extracted Data Set

It is the chemicals involved in the processing which could be increasing the risk of cancer. High temperature cooking, such as on a barbecue, can also create carcinogenic chemicals.

In the UK, around six out of every 100 people get bowel cancer at some point in their lives.

If they were all had an extra 50g of bacon a day for the rest of their lives then the risk would increase by 18% to around seven in 100 people getting bowel cancer.

So that's one extra case of bowel cancer in all those 100 lifetime bacon-eaters, argued Sir David Spiegelhalter, a risk professor from the University of Cambridge.

How bad?

The WHO has come to the conclusion on the advice of its International Agency for Research on Cancer, which assesses the best available scientific evidence.

It has now placed processed meat in the same category as plutonium, but also alcohol as they definitely do cause cancer.

However, this does not mean they are equally dangerous. A bacon sandwich is not as bad as smoking.

For an individual, the risk of developing colorectal (bowel) cancer because of their consumption of processed meat remains small, but this risk increases with the amount of meat consumed, Dr Kurt Straif from the WHO said.

Media caption Is processed meat going to kill me?

Estimates suggest 34,000 deaths from cancer every year could be down to diets high in processed meat.

Red meat risk

A.1.2 Key Terms Identification

Table A-2: Extracted Key Terms in Experiment 1

Extracted Key Terms	
Web page No. 1	Germany - Syrian - passports - asylum - September - Image - copyright - AFP - claimants - German - customs officers - seized - packages - police - illegally - asylum seekers - finance ministry - official - genuine - forged - packets - intercepted - post - letting - register - entered - EU - As - refugees - Syrian civil war- The - declined - customs - checks - German police - investigating - The EU - border - agency - Frontex - trafficking - Turkey- Fabrice - French - radio station - Europe 1 - speak - Arabic- They - North Africa - Middle East - profile - economic migrants - number - applicants - Afghans- western Balkan - Balkan countries - major - transit - country - heading - housing - camps.
Web page No. 2	Hajj - official - Saudi - toll - October - Middle East - Image - copyright - AP - Pilgrims - crushed - large - met - crush - Mecca -killed - times - Saudi Arabia - tally - Associated Press - Monday - tragedy - pilgrimage -

Table A-2: Extracted Key Terms in Experiment 1

	The - figure - media - reports - lost - The crush - incident - strike - updated - September - Iran - worst - nation - Many - dead - Africa - Nigeria - Mali - Egypt count - The AP - late - Sunday night - According - press agency - SPA - Crown Prince Mohammed - bin - Naif bin Abdul Aziz - interior minister - An - investigation - ordered - King - Salman - crown prince - progress - investigations - report
Web page No. 3	Processed meats - cancer - October - Image - copyright - sausages - ham - World Health Organization (WHO) - report - day - slices - developing - colorectal cancer - Meanwhile - red - meats - carcinogenic - limited - evidence - The WHO - stress - health - Cancer Research UK - reason - cut - occasional - sandwich - harm - What - modified - shelf life - change - methods - smoking - curing - adding - salt - preservatives - beef - mince - processed - hot dogs - salami - corned beef - beef jerky - canned - sauces - It - chemicals - involved - processing - High - temperature - barbeque - create - In - UK - point - lives - If - extra - rest - So - case - bowel cancer - lifetime - argued - Sir - David Spiegelhalter - professor - University - Cambridge - How - bad - advice - International Agency -Research - scientific evidence - plutonium - alcohol - However - equally - dangerous - individual - colorectal - (bowel) - consumption - consumed - Dr - Kurt - Media - kill - Estimates - diets -Red meat

Table A-3: Extracted Key Terms in Experiment 2

Extracted Key Terms	
Web page No. 1	Germany - Syrian - asylum - AFP - German - asylum seekers - packets - EU - Syrian civil war - German - Frontex - Turkey - French - radio station - Europe 1 - Arabic - North Africa - Middle East
Web page No. 2	Hajj - Middle East - AP - Pilgrims - Mecca - Saudi Arabia - Associated - Press - Iran - Africa - Nigeria - Mali - Egypt - Naif bin Abdul Aziz - crown prince
Web page No. 3	Processed meats - cancer - sausages - World Health Organization (WHO) - colorectal cancer - red - carcinogenic - The WHO - Cancer Research UK - sandwich - shelf life - smoking - salt - preservatives - beef - sausages - hot dogs - salami - corned beef - beef jerky - ham - UK - bowel cancer - David Spiegelhalter - Cambridge - plutonium - alcohol - bowel

Table A-4: Extracted Key Terms in Experiment 3

Extracted Key Terms	
Web page No. 1	Germany – Syrian - asylum seekers - Syrian civil war – German – Frontex – French - radio station - Europe 1 – Arabic - North Africa - Middle East – Turkey
Web page No. 2	Hajj - Middle East – Mecca - Saudi Arabia - Associated Press – Iran – Nigeria – Mali – Egypt - Naif bin Abdul Aziz - crown prince
Web page No. 3	Processed meats – cancer - sausages - World Health Organization (WHO) - colorectal cancer – red – carcinogenic - The WHO - Cancer Research UK – sandwich – salt – preservatives - hot dogs – salami - corned beef - beef jerky – carcinogenic - bowel cancer - David Spiegelhalter – Cambridge – plutonium – sandwich – bowel

Table A-5: Extracted Key Terms in Experiment 4

Extracted Key Terms	
Web page No. 1	Germany - Syrian civil war – Frontex – Turkey - Europe 1 – Arabic - North Africa
Web page No. 2	Hajj – Mecca - Saudi Arabia - Associated Press - Iran – Nigeria – Mali – Egypt - Naif bin Abdul Aziz - crown prince
Web page No. 3	Processed meats - World Health Organization (WHO) – colorectal – cancer – red – carcinogenic - The WHO - Cancer Research UK – salt – salami - corned beef - beef jerky - bowel cancer - David Spiegelhalter – Cambridge – plutonium – bowel

Table A-6: Extracted Key Terms in Our proposed approach

Extracted Key Terms	
Web page No. 1	Germany – Syrian - asylum seekers - Syrian civil war – German – Frontex – French - radio station - Europe 1 – Arabic - North Africa - Middle East – Turkey – EU – AFP – packets

Table A-6: Extracted Key Terms in Our proposed approach

Web page No. 2	Hajj – Mecca - Saudi Arabia - Associated Press – Iran – Nigeria – Mali – Egypt - Naif bin Abdul Aziz - crown prince – Africa - Middle East – AP
Web page No. 3	Shelf life – Alcohol – Cambridge – Meat – The Who – Salt – Hot dog – Salami – World Health Organization- preservatives – Beef – Sausage – Jerky – Cancer Research UK – Corned beef – Colorectal cancer – endometrial cancer – red – Sandwich – Smoking – David Spiegelhalter – Carcinogen – Ham – Plutonium – United Kingdom – Cancer

A.1.3 Semantic Annotation Accuracy:

Table A-7: Informative Annotation Process Accuracy in Experiment 1

Information Retrieval Accuracy				
	Key Term	User 1	User 2	User 3
First Data Set	Germany	T	T	T
	Syrian	T	T	T
	passports	T	T	T
	asylum	T	T	T
	September	T	T	T
	Image	T	T	T
	copyright	T	T	T
	AFP	F	T	T
	claimants	F	F	F
	German	F	F	F
	customs officers	F	F	F
	seized	T	F	F
	packages	T	F	F
	police	T	T	T
	illegally	T	F	F
	asylum seekers.	T	F	F
	finance ministry	F	T	T

Table A-7: Informative Annotation Process Accuracy in Experiment 1

official	F	F	F
genuine	F	F	F
forged	T	F	F
packets	F	F	F
intercepted	F	F	F
post.	T	F	F
letting	F	F	F
register	F	F	F
entered	F	F	F
EU.	T	T	T
As	F	F	F
refugees	T	T	T
Syrian civil war	T	T	T
The	F	F	F
declined	F	F	F
customs	T	T	T
checks.	F	F	F
German police	T	F	F
investigating.	T	F	F
The EU	T	T	T
border	F	T	T
agency	F	F	F
Frontex	T	T	T
trafficking	T	F	F
Turkey	T	T	T
Fabrice	F	F	F
French	T	T	T
radio station	T	F	F
Europe 1	T	T	T
speak	T	T	T
Arabic	T	T	T

Table A-7: Informative Annotation Process Accuracy in Experiment 1

	They	F	F	F
	North Africa,	T	T	T
	Middle East,	T	T	T
	profile	F	F	F
	economic migrants	T	F	F
	number	F	F	F
	applicants	F	F	F
	Afghans,	T	T	T
	western Balkan	T	F	F
	Balkan countries.	T	T	T
	major	F	T	T
	transit	T	T	T
	country	T	F	F
	heading	F	F	F
	housing	F	T	T
	camps.	T	T	T
Second Data Set	Hajj	T	T	T
	official	F	F	F
	Saudi	F	F	F
	toll	F	F	F
	October	T	T	T
	Middle East	T	T	T
	Image	T	T	T
	copyright	T	T	T
	AP	T	T	T
	Pilgrims	T	T	T
	crushed	T	T	T
	large	F	F	F
	met	F	F	F
	crush	T	T	T
	Mecca	T	T	T

Table A-7: Informative Annotation Process Accuracy in Experiment 1

killed	F	F	F
times	F	F	F
Saudi Arabia	T	T	T
tally	T	T	T
Associated Press	T	T	T
Monday	F	F	F
tragedy	T	T	T
pilgrimage	T	T	T
The	F	F	F
figure	F	F	F
media	F	F	F
reports	F	F	F
lost	F	F	F
The crush	F	F	F
incident	F	F	F
strike	F	F	F
updated	F	F	F
September.	T	T	T
Iran	T	T	T
worst	F	F	F
nation.	F	F	F
Many	F	F	F
dead	T	T	T
Africa	T	T	T
Nigeria	T	T	T
Mali	T	T	T
Egypt	T	T	T
count	F	F	F
The AP	T	T	T
late	F	F	F
Sunday night.	F	F	F

Table A-7: Informative Annotation Process Accuracy in Experiment 1

	Accordinging	F	F	F
	press agency,	T	T	T
	SPA,	F	F	F
	Crown Prince Mohammed	T	T	T
	bin	F	F	F
	Naif bin Abdul Aziz,	T	T	T
	interior minister,	T	T	T
	An	F	F	F
	investigation	T	T	T
	ordered	F	F	F
	King	T	T	T
	Salman	F	F	F
	crown prince	T	T	T
	progress	F	F	F
	investigations	F	F	F
	report	F	F	F
Third Data Set	Processed meats	F	T	T
	cancer -	T	T	T
	October	F	F	F
	Image	T	T	T
	copyright	T	T	T
	sausages	T	T	T
	ham	T	T	T
	World Health Organization (WHO).	T	T	T
	report	T	T	T
	day	F	T	T
	slices	F	F	F
	developing	F	F	F
	colorectal cancer	T	T	T
	Meanwhile,	F	T	T
	red	F	F	F

Table A-7: Informative Annotation Process Accuracy in Experiment 1

meats	T	T	T
carcinogenic	T	T	T
limited	F	F	F
evidence.	T	F	F
The WHO	F	T	T
stress	F	T	T
health	T	T	T
Cancer Research UK	T	T	T
reason	T	T	T
cut	F	T	T
occasional	F	F	F
sandwich	T	T	T
harm.	F	T	T
What	F	F	F
modified	F	F	F
shelf life	T	T	T
change	F	F	F
methods	T	F	F
smoking,	T	F	F
curing,	T	F	F
adding	T	T	T
salt	T	T	T
preservatives.	T	T	T
beef	T	T	T
mince	T	F	F
processed	F	F	F
hot dogs,	T	T	T
salami,	T	T	T
corned beef,	T	T	T
beef jerky	T	T	T
canned	T	T	T

Table A-7: Informative Annotation Process Accuracy in Experiment 1

sauces.	T	T	T
It	F	F	F
chemicals	T	T	T
involved	F	F	F
processing	T	T	T
High	F	F	F
temperature	T	T	T
barbeque,	T	T	T
create	F	F	F
In	F	F	F
UK,	T	T	T
point	F	T	T
lives.	T	T	T
If	F	F	F
extra	F	F	F
rest	F	T	T
So	F	F	F
case	F	F	F
bowel cancer	T	T	T
lifetime	F	F	F
argued	T	F	F
Sir	F	F	F
David Spiegelhalter	T	T	T
professor	T	T	T
University	T	T	T
Cambridge.	T	T	T
How	F	F	F
bad	F	F	F
advice	F	F	F
International Agency	T	F	F
Research	T	F	F

Table A-7: Informative Annotation Process Accuracy in Experiment 1

	scientific evidence.	T	F	F
	plutonium	T	T	T
	alcohol	T	T	T
	However,	F	F	F
	equally	F	F	F
	dangerous.	T	F	F
	individual,	T	T	T
	colorectal	T	T	T
	(bowel)	T	F	F
	consumption	T	F	F
	consumed,"	F	F	F
	Dr	T	F	F
	Kurt	F	F	F
	Media	T	F	F
	kill	T	T	T
	Estimates	F	T	T
	diets	T	T	T
	Red meat	T	T	T

Table A-8: Informative Annotation Process Accuracy in Experiment 2

Information Retrieval Accuracy				
	Key Term	User 1	User 2	User 3
First Data Set	Germany	T	T	T
	Syrian	T	T	T
	asylum	T	F	F
	AFP	F	T	T
	German	T	T	T
	asylum seekers.	T	T	T
	packets	T	F	F
	EU.	T	T	T
	Syrian civil war,	T	T	T

Table A-8: Informative Annotation Process Accuracy in Experiment 2

	German	T	T	T
	Frontex	T	T	T
	Turkey.	T	T	T
	French	T	T	T
	radio station	T	F	F
	Europe 1	T	T	T
	Arabic.	T	T	T
	North Africa	T	T	T
	Middle East,	T	T	T
Second Data Set	Hajj	T	T	T
	Middle East	T	T	T
	AP	T	T	T
	Pilgrims	T	F	F
	Mecca	T	T	T
	Saudi Arabia	T	T	T
	Associated	T	T	T
	Press	T	T	T
	Iran	T	T	T
	Africa	T	T	T
	Nigeria	T	T	T
	Mali	T	T	T
	Egypt	T	T	T
Naif bin Abdul Aziz	T	T	T	
crown prince	T	T	T	
Third Data Set	Processed meats	F	T	T
	cancer	T	T	T
	sausages	T	T	T
	World Health Organization (WHO).	T	T	T
	colorectal cancer	T	T	T
	red	F	F	F
	carcinogenic	F	T	T

Table A-8: Informative Annotation Process Accuracy in Experiment 2

	The WHO	F	T	T
	Cancer Research UK	F	T	T
	sandwich	T	T	T
	shelf life	T	T	T
	smoking,	T	T	T
	salt	T	T	T
	preservatives.	T	T	T
	beef	T	T	T
	sausages,	T	T	T
	hot dogs,	T	T	T
	salami,	T	T	T
	corned beef,	T	T	T
	beef jerky	T	T	T
	ham	T	T	T
	UK,	T	T	T
	bowel cancer.	T	T	T
	David Spiegelhalter,	T	T	T
	Cambridge.	T	T	T
	plutonium	T	T	T
	alcohol	T	T	T
	Bowel	T	T	T

Table A-9 Informative Annotation Process Accuracy in Experiment 3

Information Retrieval Accuracy				
	Key Term	User 1	User 2	User 3
First Data Set	Germany	T	T	T
	Syrian	T	T	T
	asylum seekers.	T	F	F
	Syrian civil war,	T	T	T
	German	T	T	T

Table A-9 Informative Annotation Process Accuracy in Experiment 3

	Frontex	T	T	T
	French	T	T	T
	radio station	T	T	T
	Europe 1	T	F	F
	Arabic.	T	T	T
	North Africa,	T	T	T
	Middle East	T	T	T
	Turkey	T	T	T
Second Data Set	Hajj	T	T	T
	Middle East	T	T	T
	Mecca	T	T	T
	Saudi Arabia	T	T	T
	Associated Press	T	T	T
	Iran	T	T	T
	Nigeria	T	T	T
	Mali	T	T	T
	Egypt	T	T	T
	Naif bin Abdul Aziz	T	T	T
	crown prince	F	T	F
Third Data Set	Processed meats	T	F	F
	cancer	T	T	T
	sausages	T	T	T
	World Health Organization (WHO).	T	T	T
	colorectal cancer	T	F	F
	red	F	F	F
	carcinogenic"	T	F	F
	The WHO	F	F	F
	Cancer Research UK	T	T	T
	sandwich	T	T	T
	salt	T	T	T
preservatives.	T	T	T	

Table A-9 Informative Annotation Process Accuracy in Experiment 3

	hot dogs,	T	T	T
	salami,	T	T	T
	corned beef,	T	T	T
	beef jerky	T	T	T
	carcinogenic	T	T	T
	bowel cancer	T	T	T
	David Spiegelhalter	T	T	T
	Cambridge.	T	T	T
	The WHO	F	T	T
	plutonium,	T	T	T
	sandwich	T	T	T

Table A-10: Informative Annotation Process Accuracy in Experiment 4

Information Retrieval Accuracy				
	Key Term	User 1	User 2	User 3
First Data Set	Germany	T	T	T
	Syrian civil war,	T	T	T
	Frontex	T	T	T
	Turkey.	T	T	T
	Europe 1	T	F	F
	Arabic.	T	T	T
	North Africa	T	T	T
	Hajj	T	T	T
Second Data Set	Mecca	T	T	T
	Saudi Arabia	T	T	T
	Associated Press	T	T	T
	Iran	T	T	T
	Nigeria	T	T	T
	Mali	T	T	T
	Egypt	T	T	T

Table A-10: Informative Annotation Process Accuracy in Experiment 4

	Naif bin Abdul Aziz	T	T	T
	crown prince	T	F	F
Third Data Set	Processed meats	F	T	T
	World Health Organization (WHO).	T	T	T
	colorectal cancer	T	T	T
	red	F	F	F
	carcinogenic"	T	T	T
	The WHO	F	F	F
	Cancer Research UK	T	T	T
	salt	T	T	T
	salami,	T	T	T
	corned beef,	T	T	T
	beef jerky	T	T	T
	bowel cancer	T	T	T
	David Spiegelhalter	T	T	T
	Cambridge.	T	T	T
	plutonium	T	T	T
(bowel)	T	T	T	

Table A-11: Informative Annotation Process Accuracy in Our proposed approach

Information Retrieval Accuracy				
	Key Term	User 1	User 2	User 3
First Data Set	Germany	T	T	T
	Syrian	T	T	T
	asylum seekers	T	T	T
	Syrian civil war	T	T	T
	German	T	T	T
	Frontex	T	T	T
	French	T	T	T
	radio station	T	T	T

Table A-11: Informative Annotation Process Accuracy in Our proposed approach

	Europe 1	T	T	T
	Arabic	T	T	T
	North Africa	T	T	T
	Middle East	T	T	T
	Turkey	T	T	T
	EU	T	T	T
	AFP	F	F	F
	Refugees	T	T	T
	Packets	F	F	F
Second Data Set	Hajj	T	T	T
	Mecca	T	T	T
	Saudi Arabia	T	T	T
	Associated Press	T	T	T
	Iran	T	T	T
	Nigeria	T	T	T
	Mali	T	T	T
	Egypt	T	T	T
	Naif bin Abdul Aziz	T	T	T
	crown prince	T	T	T
	Africa	T	T	T
	Middle East	T	T	T
	AP	T	T	T
Third Data Set	Shelf life	T	T	T
	Alcohol	T	T	T
	Cambridge	T	T	T
	Meat	T	T	T
	The Who	F	F	F
	Salt	T	T	T
	Hot dog	T	T	T
	Salami	T	T	T
World Health Organization-	T	T	T	

Table A-11: Informative Annotation Process Accuracy in Our proposed approach

	preservatives	T	T	T
	Beef	T	T	T
	Sausage	T	T	T
	Jerky	T	T	T
	Cancer Research UK	T	T	T
	Corned beef	T	T	T
	Colorectal cancer	T	T	T
	endometrial cancer	T	T	T
	red	F	F	F
	Sandwich	T	T	T
	Smoking	T	T	T
	David Spiegelhalter	T	T	T
	Carcinogen	T	T	T
	Ham	T	T	T
	Plutonium	T	T	T
	United Kingdom	T	T	T
	Cancer	T	T	T