ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse116862019.pdf
https://doi.org/10.30534/ijatcse/2019/116862019

Toward a Semantic search engine for E-Commerce

Abdelhadi BAHAFID[1], Kamal El GUEMMAT[2], [3]El Habib BEN LAHMAR [3,] Mohamed TALEA
[1]Laboratory of Information Processing, Faculty of Sciences Ben M'Sik, Hassan II University, Casablanca, Morocco, bahafidabdelhadi@gmail.com
[2]Laboratory of Information Processing, Faculty of Sciences Ben M'Sik, Hassan II University, Casablanca, Morocco, k.elguemmat@gmail.com
[3]Modeling and Information Technology Laboratory, Faculty of Sciences Ben M'Sik, Hassan II University, Casablanca, Morocco, h.benlahmer@gmail.com
[4]Laboratory of Information Processing, Faculty of Sciences Ben M'Sik, Hassan II University, Casablanca, Morocco, taleamohamed@yahoo.fr

ABSTRACT

E-commerce occupies an important place in the user's lives, it makes their lives easier. Allows users to find products, using search engines, which to respond to users, browse all e-commerce web pages, extract and index the information displayed on these pages. This extraction is done in a classic way by the search engine like any other non-e-commerce page, because it does not recognize the properties of the product by parsing the page, it can't recognize the brand, the type of product ..., it extracts the text and indexes the raw text and the properties of the product are presented as raw HTML.

The main problem of e-Commerce is the extraction of relevant information about products and this is because the information is presented in a way that is difficult for machines to understand, which can be explained by the rarity of use of Semantic Web technologies, and the lack of proper standards where HTML does not provide the syntax and semantics of information.

Semantic Web Technologies enable machines to interpret data published in a machine-interpretable form on the web, which facilitates the extraction, processing and indexing of this information. Now, only human beings can understand the product information published online. The emerging semantic Web technologies have the potential to deeply influence the further development of the Internet Economy.

These different problems cited, the scarcity of semantic search engines and the lack of work on this subject pushed us to deepen a bit of our research and try to find more works of e-commerce and semantic web, study and analyze them to have an idea about the progress of this area.

The purpose of this article is to present an overview of the semantic web technologies used in the field of e-commerce, the different limitations of the current e-commerce, the impact of the use of these technologies in e-commerce, present some of the most used e-commerce search engines and propose a new approach to create a new semantic search engine, that can be used by the community to achieve our goal of creating a first semantic search engine for e-commerce.

Key words: E-commerce, Search engine, Semantic indexing, Ontology, Web semantic.

1. INTRODUCTION

The web changes, we went from the static web of the origins (1.0) to the participative web (2.0), then to the semantic web (3.0) [25] which allows machines to understand the meaning of the data and to better exploit them. This progression has not been adopted by all areas and its implementation is a bit slow.

By "semantics", that does not mean that the machine understands in the same way as humans the information contained in each of these pages. However, this information (data) can be the subject of a structured language describing this data, and sufficiently standardized to be shareable by machines. This language is called "metadata" (data describing data, or metadata). In the world of documents, such "metadata" have existed for a long time. In libraries, bibliographic records of documents contain structured information describing for example a book - Author = Victor Hugo, Title = "Les Misérables", etc. Author, Title, Publisher, Date, etc. are all standardized metadata in an exchange format that allows all library systems to share and process this information (distinguish editions for example, manage loans, acquisitions, etc.). We note that metadata are now often produced at the same time as the data, for example for photos (format, date, geolocation, pattern recognition, color, etc.), with digital cameras, for which we can talk about "embedded" metadata.

For e-commerce we notice that the product catalog is the wealth of the e-Commerce and the website is its materialization, the "virtual" showcase for users. This is why

its maintenance, enrichment and valuation are key success factors for the merchant sites. The complexity of this task is up to the challenge and this work can be quickly tedious and cumbersome to improve the visibility of products on mainstream search engines, the natural referencing techniques [9] being more and more complex to maintain with the increase of the competition, than to expose the data of the catalog with the partners who each present particularities in the absence of normalization or for the aggregation heterogeneous data from different providers. The lack of html standard to present the heterogeneous e-commerce data in the same way also complicates robot's data extraction tasks which explains the poor quality of search engine results, which forces users to spend a lot of time sorting, comparing products between sites and choosing the right product that suits their needs.

All problems cited among which the lack of standard HTML allowing to express the semantics of the various attributes of the product, which implies that the web pages are indexed in a traditional way, the absence of semantic e-commerce search engine which explains the low proportions of the recall and precision, maintenance, enrichment and valuation of product catalogue, spend a lot of time looking for products that meet user needs and the scarcity of the works carried out in this field, not enough research and tries to improve the quality of the e-commerce service, pushed us to study some works done especially those that deal with all that is semantic search, tools used to implement the semantic web in e-commerce to highlight the existing and the different possibilities to improve it.

Faced with these difficulties, semantic Web technologies [6] now offer solutions. The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from many researchers and industrial partners.

The result of this work is a study of different technologies most used in the semantic web for e-commerce and a new approach proposed in the section "Toward a new semantic search engine" to create the first semantic search engine for e-commerce, using semantic extraction method proposed in other article [4] and GoodRelations ontology, this search engine will make life easier for users when searching for different products on the internet.

This article is organized as follows: Section 2 presents an overview of semantic web technologies that allow producers to describe their resources (products) to facilitate users' recovery of targeted products, Section 3 describes the current e-commerce, its limitations, section 4 the contribution of semantic web tools on e-commerce made available and the presentation of some e-commerce search engines, section 5 describe an approach to create an ontology based semantic search engine, section 6 result and discussion about the approach and section 7 conclusion and perspective of this paper.

2. THE SEMANTIC WEB TECHNOLOGIES

The Semantic Web is a vision about an extension of the existing World Wide Web, which provides software programs with machine-interpretable metadata of the published information and data. In other words, we add further data descriptors to otherwise existing content and data on the Web. As a result, computers can make meaningful interpretations like the way human process information to achieve their goals.

The ultimate ambition of the Semantic Web, as its founder Tim Berners-Lee sees it, is to enable computers to better manipulate information on our behalf. He further explains that, in the context of the Semantic Web, the word "semantic" indicates machine-processable or what a machine can do with the data [21]. Whereas "web" conveys the idea of a navigable space of interconnected objects with mappings from URIs to resources.

Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Those technologies such as RDF, OWL, SPARQL...

2.1. RDF

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en">
    <head>
        <title>Web semantic</title>
    </head>
    <body>
        <div id="header">
            <ul id="menu">
                <li><a href="/">Home page</a></li>
                <li><a href="/contact/">Contact</a></li>
            </ul>
        </div>
        <h1>towards a semantic search engine</h1>
        <h2>Semantic</h2>
        <p>
            By "semantics", that does not mean that the machine understands in the same way as humans the information contained in each of these pages.
            However, this information (data) can be the subject of a structured language describing this data, and sufficiently standardized to be
            shareable by machines.

        </p>
        <p>
            The content of this pages is the sole responsibility of the authors of these sites <a href="/licence/">More details.</a>
        </p>
    </body>
</html>
```

Figure 1: RDF Example

If a user reads this part of code, he can distinguish what is a title, who is the author, what is the menu of the site, what is the license of this page.

The person is interested only in the content of the page reading the text and the rest it doesn't matter too much, while the search engine to perform search by author or license used

by this page we must explicitly indicate for it this information!

Finally, let's compare what the browser sees and what the human sees. The browser sees a list, a title, a lower level title and paragraphs. It does not make sense to it. The human sees that there is a menu, which will help him navigate the site, he sees the page with title that this page was written by someone whose name is written, that this page is under license. The browser should also have access to this information.

To give this additional information, we must annotate the content of the HTML pages (Figure 1). We will use here RDFa (is a serialization of RDF).

Resource Description Framework (RDF) [26] is a standard, a W3C recommendation that represents a graph model for describing Web resources and their metadata, to allow automatic processing of such descriptions. It was developed by the W3C, RDF is the basic language of the Semantic Web. One of the syntaxes (or serializations) of this language is RDF RDFa. Other RDF syntaxes appeared, seeking to make reading of resources more understandable, this is the case for example of Notation3 (or N3).

RDF extends the linking structure of the Web to use URIs to name the relationship between objects as well as both ends of the link (what is usually called a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed and shared among different applications.

An RDF structured document is a set of triplets [28], an RDF triplet is an association (subject, predicate, and object):

- The "subject" represents the resource to be described identified by a URI.
- The "predicate" represents a type of property or binary relation on the domain between the subject and the result of this predicate object applicable to this resource.
- The "object" represents a data or other resource: it is the value of the property.

2.2. RDFa

RDFa [29] is a syntax for describing structured data in a web page. The RDFa code is invisible to the user and does not affect the content of a web page. RDFa, provides a set of XHTML attributes [1] to enrich visual data with machine-readable information.

RDFa is a set of elements and attributes (example in Figure 2).

```
<div typeof="foaf:Person" xmlns:foaf="http://xmlns.com/foaf/0.1/">
   <p property="foaf:name">
      abdelhadi BAHAFID
   </p>
   <p>
      Email: <a rel="foaf:mbox" href="mailto:abdelhadi@semantic.com">abdelhadi@semantic.com</a>
   </p>
   <p>
      Phone: <a rel="foaf:phone" href="tel:+33-6-55-57-33-44">+33655733344</a>
   </p>
</div>
```

Figure 2: RDFa usage example

2.3. RDFS

RDF Schema (RDFS) is extending RDF vocabulary to allow describing taxonomies of classes and properties. It also extends definitions for some of the elements of RDF, for example it sets the domain and range of properties and relates the RDF classes and properties into taxonomies using the RDFS vocabulary, is a set of classes with certain properties using the RDF extensible knowledge representation data model, providing basic elements for the description of ontologies, otherwise called RDF vocabularies, intended to structure RDF resources. These resources can be saved in a triple store to reach them with the query language SPARQL.

The RDF Schema class and property system is similar to the type systems of object-oriented programming languages such as Java. RDF Schema differs from many such systems in that instead of defining a class in terms of the properties its instances may have, RDF Schema describes properties in terms of the classes of resource to which they apply. This is the role of the domain and range mechanisms described in this specification.

Above an example RDFs (Figure 3)

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"xmlns:dc="http://purl.org/dc/elements/1.1/">

<rdf:Description rdf:about="http://www.exemple.com/websem.pdf">
   <dc:creator>Abdelhadi BAHAFID</dc:creator>
   <dc:title>web of data tutorial</dc:title>
      <dc:description>
      Linked open data is linked data that is open data. Tim Berners-Lee gives the clearest definition of linked open data in
      differentiation with linked data.
      </dc:description>

   <dc:subject>
      <rdf:Bag>
         <rdf:li>Web of data</rdf:li>
         <rdf:li>RDF</rdf:li>
         <rdf:li>RDF-Schema</rdf:li>
      </rdf:Bag>
   </dc:subject>
   <dc:date>2018-02-01</dc:date>
</rdf:Description>
</rdf:RDF>
```

Figure 3: RDFS example

2.4. LIMITATIONS RDF/RDFS

RDF and RDFS allow to define data or metadata as graphs of triplets. However, many limitations [8] limit the ability to express knowledge established using RDF / RDFS. For example:

• The inability to reason and carry out automated reasoning on RDF / RDFS knowledge models.
• RDF-S does not allow to express that 2 classes are disjoint
• RDF-S does not allow to set a restriction on the number of occurrences of values that a property can take
• RDF-S does not allow to characterize properties including:
  - transitivity: eg: isMoreBigThan
  - uniqueness: Ex: isFatherOf
  - inverse property: Ex: "eat" = inverse property of "isEatenBy"

It is this lack that OWL intends to fill.

2.5. OWL (Ontology Web Language)

Web Ontology Language (OWL) [19] owes its name to the term "ontology", a word borrowed from philosophy, it is of Greek origin, was obviously created only in the 17th century. According to Aristotle "Speech on being as being", ontology takes completely different meaning in computer science, where the term refers to a structured set of knowledge in a field of knowledge.

Web Ontology Language (OWL) is a language for representing rich and complex knowledge about objects, groups of objects, and relationships between them. OWL is a logic-based computation language such that the knowledge expressed in OWL can be exploited by computer programs, for example to check the consistency of this knowledge or to make knowledge explicit or implicit. OWL is a logic-based computation language such that the knowledge expressed in OWL can be exploited by computer programs, for example to check the consistency of this knowledge or to make implicit knowledge explicit. OWL documents, called ontologies, can be published on the Web and can refer to other OWL ontologies. OWL [7] is part of the technological stack of the W3C Semantic Web, which includes RDF, RDFS, SPARQL, etc.

A. Why OWL

The Semantic Web is a vision of the future Web, where the information gets an explicit meaning facilitating the automatic processing and integration of information available on the Web by machines. The Semantic Web will be built with XML [10][31] and its ability to define custom markup schemes, and with RDF its flexibility to represent data. An ontological language, able to formally describe the meaning of the terminology used in Web documents, constitutes the first necessary level of the Semantic Web after RDF. If we want machines to perform useful reasoning tasks on documents, then the language must go beyond the basic semantics of the RDF schema. OWL provides more details on ontologies which justifies the need for a Web ontology language.

In practice, OWL is designed as an extension [7] of RDF and RDF Schema (RDFS). OWL is designed for class descriptions (by constructors) and property types. Therefore, it is more expressive than RDFS, to which some people criticize a lack of expressiveness due to the unique definition of the relations between objects by assertions. OWL also brings better integration, evolution, sharing and easier inference of ontologies.

OWL provides three sub-languages of expression for specific developer and user communities: OWL Lite [25], The OWL DL language [19] and The OWL Full [5].

2.6. SPARQL

RDF is a graphical data format that is oriented and labeled to represent information in the Web. SPARQL [23] can be used to express queries across various data sources [32], whether the data is stored natively as RDF or seen as RDF via middleware. SPARQL can search for mandatory and optional graph patterns [22] as well as their conjunctions and disjunctions.

It defines also extensible value testing and expression framework. It presents the functions and operators that can be used to constrain the values that appear in a query's results and also calculate new values to be returned by a query.

SPARQL is the query language of the Semantic Web. This allows us:
- Extract structured and semi-structured data values
- Explore the data by querying unknown relationships
- Perform complex database joins in a single simple query
- Transform RDF data from a vocabulary to another

3. E-COMMERCE

Electronic Commerce or eCommerce can be defined as the exchange of goods and services by means of the Internet (Web) or other computer network infrastructures. eCommerce follows the same basic principles as traditional commerce—that is, buyers and sellers come together to exchange goods for money. In eCommerce, buyers and sellers transact business over networked computers, which can be across cities, countries or continents. There are two major eCommerce styles, they are: Business-2-Consumer and Business-2-Business eCommerce models.

The Web is moving from a collection of pages towards an accumulation of services [2] that interoperate through the Internet. Consequently, people Worldwide are carrying out more and more commercial activities rather than simple information readings on the Web.

However, current ecommerce is experiencing many restrictions. On the one hand, ecommerce is asked to perform more intelligently and autonomously for fitting the changing situation. On the other hand, due to lack of a meaningful (semantic) description, machines are unable to handle the ecommerce tasks sophisticatedly in the current Web context.

3.1. Current Status of e-commerce

The internet has changed the way people live so far. Online services have enabled people from all walks of life to bring entire libraries, entertainment venues, post offices and financial centers to a workplace, office or shirt pocket. The biggest and most significant impact of the Internet may lie in the way consumers buy every-thing from gifts to gadgets to shopping, clothing, cars and cruises. All these business activities, including online shopping and online banking, make up the concept of e-commerce.

A search for any product offers is the starting point for most eCommerce transactions. ECommerce web applications are designed to return the most appropriate data to the user based on limited keywords supplied by the user, and the current applications are failing in returning the relevant data to the consumers. Many limitations are observed on current ecommerce models.

3.2. Limitations

Today, eCommerce greatly changed our lifestyle. However, with the further investigation, conventional eCommerce technology has been found several problems and limitations. On the one hand, the quantity of eCommerce increases faster and faster. On the other hand, the quality of eCommerce appears to be information asymmetric. The same product may have several providers at different prices. For the example of the case: the Samsung S9 phone. There are several Web sites that sell the exactly same product at different prices. In Cdiscount, the product is sold for 599 euros, however, the same product being sold at Boulanger.fr has a cost 709 euros, total saving is 200 euros.

A search for any product or product offers is the starting point for most e-Commerce transactions. E-Commerce web applications are designed to return the most appropriate data to the user, but the current applications are failing in returning the relevant data to the consumers. Following limitations are observed on current ecommerce [20]:

a. Information Asymmetry & Price Dispersion: This situation occurs where the same product with same features is available with different price values in different websites to the Consumer's.

b. Semantic Description & Extension is Deficient: This situation occurs where the product's generic attributes are not considered, such as price, color, function, origin and material etc...

c. Business Attributes: This situation occurs where the customers choose the tax percentage, type of pay and discount offered if any etc.

d. Interoperability in an inconsistent environment: This situation occurs where the consumer is in the conflicting state to choose the best option from the available websites.

Hence, efficient search is the big problem facing to the traditional ecommerce.

Billions of searches are conducted every day on the Internet by people trying to find what they need. A majority of these searches are in the domain of consumer eCommerce, where a Web user is looking for something to buy. This represents a huge cost in terms of people hours and an enormous drain of resources.

Besides, the eCommerce process is getting more and more complex. Therefore, it should be carried out more intelligently and autonomously.

A typical scenario of traditional eCommerce involves a user's visiting one or several online shops, browsing their offers, selecting and ordering products. Currently these operations are manually carried out. Users should search and collect all the relevant information about prices, terms, and conditions by themselves. Obviously, this is a time-consuming and low efficient operation.

As a result, eCommerce must be carried out autonomously, with minimal human intervention. Ideally, online services should be available in the form of some descriptions. Software agents can, then, extract the product and its price information and even compile a market overview.

E-commerce must be able to function more intelligently and autonomously to adapt to changing circumstances. In the absence of a meaningful description (semantics), machines are not able to handle e-commerce tasks in a valuable way in the current web context. To this end, eCommerce must be built upon a meaningful Web infrastructure. The Semantic Web offers a favorable support for eCommerce to enrich a data with additional meaning (semantics) so that more people, objects and machines can work on it.

4. CONTRIBUTIONS OF THE SEMANTIC WEB TO ECOMMERCE

The Semantic Web is considered as the future Web, is an infrastructure that makes Web resources more accessible both for human and computers, it provides software programs with machine-interpretable metadata of the published information and data. In other words, we add further data descriptors to otherwise existing content and data on the Web. As a result, computers are able to make meaningful interpretations

similar to the human's way process information to achieve their goals [18]. The Semantic Web will give a more valuable benefit to eCommerce.

The Semantic Web brings a valuable advance to eCommerce. The limitations of current eCommerce infrastructures could be overcome by providing a semantic markup about the service description. On the one hand, the Semantic Web provides ontologies [12]. That act like shared knowledge bases across the Web. On the other hand, it also offers a logic to infer how such terms (ontologies) are combined to form complex concepts and how do they interact with the knowledge already accumulated.

Currently eCommerce is in the state of emerging for a low-level description. This description could be enriched with meaningful information using Semantic Web ontologies that act [27] for the eCommerce property and capability description, it allows to improve readability of "information or knowledge representation" and then facilitate information processing activity to identify a particular information that is contained in the text of pages and to store it in a structured form (database, XML file, OWL file,)..
Although vast amount of conceptual models, vocabularies, schemas or ontologies are available for free, we have investigated only the selected three ecommerce related standards or ontologies either fully or partially built by different research groups, product companies, practitioners and individuals. These are: UNSPSC, eClassOWL and GoodRelations ontology.

4.1. UNSPSC

The UNSPSC [11] was jointly developed by the United Nations Development Programme (UNDP) and Dun & Bradstreet in 1998 and is currently managed by GS1 US, which is responsible for overseeing code change requests, revising the codes and issuing regularly scheduled updates to the code, as well as managing special projects and initiatives.

The United Nations Standard Products and Services Code (UNSPSC) is a taxonomy of products and services for use in ecommerce. It is a four-level hierarchy coded as an eight-digit number, with an optional fifth level adding two more digits.

UNSPSC code, offers a single global classification system. It provides an open, global multi-sector standard for efficient, accurate classification of products and services.

4.2. eClassOwl

eClassOWl [15] was developed by Digital Enterprise Research Institute (DERI) University of Innsbruck originally initiated by martin hepp in 2003.eClassOWL is an OWL ontology for describing the types and properties of products and services on the Semantic Web (also known as the "Web of Linked Data"). eClassOWL is meant to be used in combination with the GoodRelations ontology for e-commerce, which covers the commercial aspects of offers and demand, e.g. prices, payment, or delivery options.

4.3. GoodRelations ontology

GoodRelations [16] is a vocabulary that can be used to exchange information about products and services, pricing, payment options, other terms and conditions, store locations and their opening hours, and many other aspects of e-commerce, between networks of computer systems. The focus is on interoperability between Web sites and clients consuming the information given on those sites. Through this vocabulary manufacturers or digital web shop assistants can describe the exact meaning of their offers.

GoodRelations is a lightweight ontology for exchanging e-commerce information, namely data about products, offers, points of sale, prices, terms and conditions, on the Web. It can be used in all RDF syntaxes (like RDF/XML, Turtle, RDFa, JSON-LD ...), Micro data, and basically any syntax that supports an Entity-Attribute-Value pattern.
GoodRelations is the most powerful vocabulary for publishing all of the details of your products and services in a way friendly to search engines, mobile applications, and browser extensions. By adding a bit of extra code to your Web content, you make sure that potential customers realize all the great features and services and the benefits of doing business with you, because their computers can extract and present this information with ease. This ontology has been developed by answering competency questions related to the location of service offers on the web, availability of services in spatial and temporal dimensions, eligibility of customers, payment options, delivery methods, and tax calculations.

The following example (Figure 4) shows that we can enrich the content of a web page with semantic information readable by the indexing robots. The information we can add:

- *"*Hepp's Bagel Bakery Ltd. ": The official legal name of your company or business.
- "Germany": Your country.
- "Munich": The city in which your business is registered.
- "85577": The zip code of your residence.
- "1234 Main Street": The street and number of your residence
- "+1 408 970-6104": The phone number, including the international prefix.
- "http://www.hepps-bagels.com/image_or_logo.png" : The Web address (URL) of a logo or image.

```
<!-- in RDFa 1.1, you can use the abbreviation:
<div profile="http://www.heppnetz.de/grprofile/">
-->
<div xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns="http://www.w3.org/1999/xhtml"
     xmlns:foaf="http://xmlns.com/foaf/0.1/"
     xmlns:gr="http://purl.org/goodrelations/v1#"
     xmlns:vcard="http://www.w3.org/2006/vcard/ns#">

  <div about="#company" typeof="gr:BusinessEntity">
    <div property="gr:legalName" content="Hepp's Bagel Bakery Ltd."></div>
    <div rel="vcard:adr">
      <div typeof="vcard:Address">
        <div property="vcard:country-name" content="Germany"></div>
        <div property="vcard:locality" content="Munich"></div>
        <div property="vcard:postal-code" content="85577"></div>
        <div property="vcard:street-address" content="1234 Main Street"></div>
      </div>
    </div>
    <div property="vcard:tel" content="+1 408 970-6104"></div>
    <div rel="foaf:depiction" resource="http://www.hepps-bagels.com/image_or_logo.png">
    </div>
    <div rel="foaf:page" resource=""></div>
  </div>
</div>
```

Figure 4: GoodRelations integration in html document

4.4. E-commerce search engine

Nowadays, the advance of Internet and Web technologies has continuously boosted the prosperity of e-commerce. Through the Internet, it has become daily life for people to online shopping, and the number of people buying, selling and performing transactions on the Web is increasing at a phenomenal pace. With the further development of e-commerce, it will not be easy for customers to single out the best commodity when faced with the massive commodity information in the Internet. Usually, customers use various E-commerce search engines to search and compare commodities when they do online shopping in the Internet. Therefore, E-commerce search engines have largely become the main methods for customer to acquire commodity in-formation and relevant services in the course of e-commerce activities.

Many search engines dedicated to e-commerce have emerged, there is a range of choices on the internet, the most used are Elasticsearch, Apache Solr, Sphinx and there are others: Searchanise, Instant Search, CloudSearch for Ecwid, Algolia, Searchly....

We will represent the most important ones.

A. Elasticsearch

Elasticsearch [13] is a distributed, RESTful search and analytics engine capable of solving a growing number of use cases, is one of the most popular search engines, used by the best ecommerce sites.

ElasticSearch allows to research any type of document. It has an adaptable architecture, searches almost in real time and can be organized in multi-entity.

B. Apache Solr

Solr (pronounced "solar") is an open source enterprise search platform, from the Apache Lucene project. Its major features include full-text search [14] hit highlighting, faceted search, real-time indexing, dynamic clustering, database integration, NoSQL features and rich document (e.g., Word, PDF) handling.

C. Sphinx

Sphinx is less popular than previous open source search servers, but it is still used by such top websites as Craigslist, Groupon, and Living Social. Sphinx is able to search both simple files and data in an SQL database or NoSQL storage. It offers lots of text processing features [3] and supports customization. Thus, you can easily adjust it according to your specific requirements.

Those search engines are keyword-based search, are not only low-efficient, but also sometimes the retrieved document contents of web pages are non-relevant with customer's query. The main reasons that result in these problems are:

1) the traditional information search techniques cannot express the semantic information correctly, and the information search based on keyword-matching still causes the semantic inaccuracy of retrieved results.
2) The heterogeneous characteristic of information organization is very obvious because of the diversity of e-commerce platform and the standard deficiency of relevant domain information description.
3) There are still not effective commodity evaluation and comparison mechanism so as to cause the information overload of the retrieval results.

4.5.Semantic E-commerce search engine

E-commerce search engines haven't invested much in understanding the semantics of a query. For queries not in the search log, the algorithms simply fall back to keyword matching. In terms of effectiveness, it often looks like the search engine tries to guess a customer's intent. Then, the customer, after being presented unsatisfactory results, tries to guess how to modify his key-word query so that the search results become more relevant. The e-commerce experience degenerates into a frustrating game of back and forth guessing.
According to our research we did not find a semantic search engine, using ontologies to present the products and facilitate the search for product information.

E-commerce must allow a smoother exchange of information and transactions between all economic actors, from the provider of products or services to end customers.

Offerers of products and services must be able to propagate and present their offers, and to the customers, to find and order the selected offer (s). By providing one-stop access to a large collection of frequently updated items or services, an e-commerce place facilitates the meeting of supply and demand through commercial mediation tools.

Ontologies achieve this goal of presenting data so that it is understandable by machines (indexing robots) and humans (users), ontology-based systems are emerging as a key technology for the development of efficient, open and profitable E-Commerce solutions. However, due to a lack of domain model and business process standards in the broader economic sectors, E-commerce is struggling to take off.
Indeed, the variety of deployed e-business and e-commerce solutions using highly diversified exchange configurations, coupled with the lack of reliability and security on the Internet, make scalability through integration and interworking impossible of these different solutions.

Moreover, in a market situation where cooperation and competition interfere, the adoption of domain standards and economic transactions is very difficult to achieve. Furthermore:

- Commercial practices are very varied and make normative alignments very difficult.
- Companies are complex: the description of products and services (alone or in combination) and their interactions are difficult to model.
- The rules of the economic game on market places are very opportunistic;
- The adoption of standards could limit commercial creativity.

Despite all these difficulties, real benefits could be derived from the use of ontologies in the following areas: categorization of products in catalogs, categorization of services (including web services), yellow pages of service companies, identification of countries, regions and currencies, identification of organizations, legal persons and entities, identification of transport containers (type, situation, road and contents) or classification of statistical data.

5. TOWARD A NEW SEMANTIC E-COMMERCE SEARCH ENGINE

The goal is to construct a semantic search engine for e-commerce that responds to client requests. To do this we will take as input a web page, from which we will extract the different information about products (information that will be indexed and used to respond to user queries). We will use an ontological approach to construct this semantic search engine.

Search engines get their information by web crawling from site to site, this task is done by Spiders. The spider extracts certain information back to be indexed depending on many factors, such as the titles, page content, JavaScript, Cascading Style Sheets (CSS), headings, or its metadata in HTML Meta tags. After a certain number of pages crawled, amount of data indexed, or time spent on the website, the spider stops crawling and moves on. No web crawler may actually crawl the entire reachable web, due to infinite websites, spider traps, spam, and other exigencies of the real web.

A search engine maintains the following processes in near real time: Web crawling, Indexing, and Searching. Our semantic search engine will maintain the following processes: Semantic Web crawling, Semantic Indexing and Semantic Searching.

5.1.Semantic Web crawling

The process of Web crawling or spidering is realized by web crawlers. A web crawler (called a web spider or web robot) is a software or automated script which parse all pages in the web automatically.

Many sites, especially search engines, use spiders to get up-to-date data. The main task of Web crawlers is to make a copy of all the visited pages, in order to be downloaded, processed and indexed by the search engine.

Once the e-commerce web pages are downloaded, we will extract the product information from these pages, using the semantic extraction approach already proposed: SEMANTIC INFORMATION EXTRACTION APPROACH FOR E-COMMERCE SEARCH ENGINE BASED ON GOODRELATIONS ONTOLOGY [4], this extraction approach make link between the embedded CSS on the e-commerce web page and the GOODRELATIONS ontology used to index these web pages through a database. The Database will contain for each site and for each attribute of a product (element) in this site the CSS class that describes it as well as the corresponding attribute in the GOODRELATIONS ontology.

Our proposed search engine, browse the sites, for each site it check if it is in its database , it means that it has all the semantic information about the site (according to the approach ([4])), For each site we generate a wrapper to extract the relevant information contained in the page, if we don't have this information, we proceed to a classic extraction and indexing of this page, the algorithm (figure 5) shows how this step takes place:

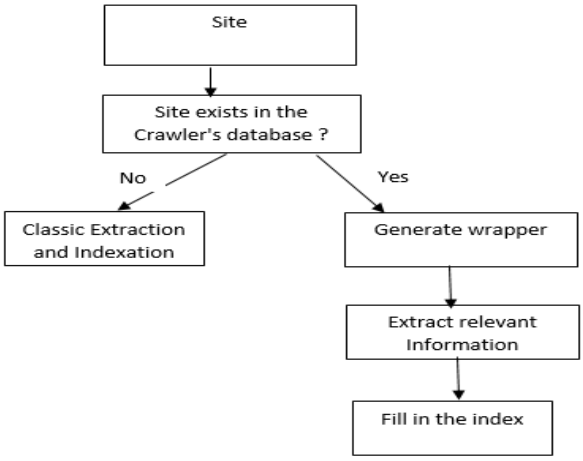

Figure 5: Algorithm of crawling sites

5.2. Semantic Indexing

The objective of this phase is to represent the different information on the product. Semantic indexing has two steps:

- Disambiguation: get the exact meaning of extracted information from the document to be indexed.
- Representation: to represent the document in order to retrieve information.

A. Disambiguation

The semantic web crawling using the semantic extraction approach: SEMANTIC INFORMATION EXTRAC-TION APPROACH FOR E-COMMERCE SEARCH ENGINE BASED ON GOODRELATIONS ONTOLOGY [4], allowed us to extract the information the different attributes of the products and know each attribute corresponds to which property (price, brand, ...) it means knowing its meaning for the product, its semantics.

B. Representation

To extract the product information, we used the GOODRELATIONs ontology, then it will be used as a model also to represent the product, since we extract we know each product attribute corresponds to which attribute of GOODRELATIONS.

5.3. Semantic Search

The user starts his search by entering keywords (user request), the request is processed and converted into a SPARQL query, this conversion is essential to be able to request our semantic repository, to do this we will use the approach "Ontology-based translation of natural language queries to SPARQL" [30] which follows two steps:

-Transforming natural language question to logical query under consideration of a lexicon and underlying KB(Knowledge Base : GoodRelations Ontology in our case)

- Transforming logical query into query under consideration of relations and topology in the KB.

The query will be run on the repository data (Index) and returns the list of documents that meet this request (figure), and the link to the relevant documents is established and the result is displayed to the user.

6. RESULT AND DISCUSSION

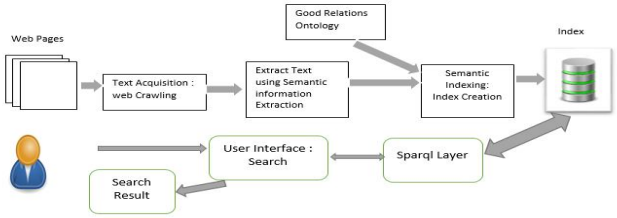

Figure 6: Proposed Semantic Search process

The proposed Semantic search (Figure 6) aims to improve the accuracy of the search by trying to understand the user's query. Through the concept mapping, the synonyms and the natural language algorithms, semantic search gives more research results than traditional keyword-based research.

6.1. Discussion

Generally, the search is keyword-based, because the indexing robots when they parse the web to extract the information from the e-commerce web pages, it extracts the information in a traditional way, there are several techniques among which weighting technique, statistics on the text… None of these techniques is semantic technique, in the end we may have an index that may contain information that has nothing to do with the content of the pages to which it refers.
Now with the semantic search engine, it parses the e-commerce page as a human being, it can know for each information, it corresponds to which attribute of the product, by means of the metadata added via GOOD-RELATION ontology which also facilitate to the robots, during the crawl of the page to extract only the relevant product information, i.e. the index will contain only relevant information, and that will improve the relevance of the search results.

7. CONCLUSION ET PERSPECTIVES

This work presents an overview of the semantic technologies used in the field of e-commerce, there are quite a few RDF, OWL, SPARQL ..., the problems that traditional e-commerce encounters, it also shows the lack of semantic search engine for e-commerce, Through an effective method of study, analysis of work done to solve the problem of lack of articles around the semantic web for e- commerce that has a great impact on the e-commerce market as well as the user lives and also propose a new approach to build a first semantic search

engine for e-commerce, who will be able to browse the e-commerce pages via semantic web technologies and extract only the relevant information on the products, since the search engine will be able to distinguish the price of the brand from the other properties, so that the answer to the user requests is more accurate with precision and recall portions more interesting.

As perspective for this work is to achieve a development of our first semantic search engine for e-commerce, based on GOODRELATIONS ontology to address all the problems and limitations mentioned above by considering a standard representation of e-commerce data and set it up for a few e-commerce sites to test its relevance and develop a new approach to convert a natural language query into a SPARQL query using an e-commerce knowledge base like GOODRELATIONS and the grammatical structure of the query.

REFERENCES

1. Adida, B., Birbeck, M., McCarron, S., & Pemberton, S. (2008). RDFa in XHTML: **Syntax and processing. Recommendation, W3C**, 7, 41. https://doi.org/10.5121/ijwest.2012.3101
2. Aghaei, S., Nematbakhsh, M. A., & Farsani, H. K. (2012). **Evolution of the world wide web: From WEB 1.0 TO WEB 4.0**. International Journal of Web & Semantic Technology, 3(1), 1
3. Aksyonoff, A. (2011). **Introduction to Search with Sphinx: From installation to relevance tuning**. " O'Reilly Media, Inc.".
4. BAHAFID A, EL GUMMAT K, BEN LEHMAR EL, TALEA, M. (2019). **SEMANTIC INFORMATION EXTRACTION APPROACH FOR E-COMMERCE SEARCH ENGINE BASED ON GOODRELATIONS ONTOLOGY**. Journal of Theoretical and Applied Information Technology, 97(2)
5. Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., & Stein, L. A. (2004). **OWL web ontology language refer-ence. W3C recommendation**, 10(02)
6. Berners-Lee, Tim, James Hendler, and Ora Lassila. "**The semantic web**." Scientific american 284.5 (2001): 28-37 https://doi.org/10.1038/scientificamerican0501-34
7. Carroll, J., Herman, I., & Patel-Schneider, P. F. (2015). **OWL 2 web ontology language RDF-based semantics**. W3C Recommendation (October 27, 2009).
8. Chernenkiy, V., Gapanyuk, Y., Nardid, A., Skvortsova, M., Gushcha, A., Fedorenko, Y., & Picking, R. (2017, September). **Using the metagraph approach for addressing RDF knowledge representation limitations**. In 2017 Internet Technologies and Applications (ITA) (pp. 47-52). IEEE.
9. Cushman, M. (2018). **Search engine optimization: What is it and why should we care?. Research and practice in thrombosis and haemostasis**, 2(2), 180-181. https://doi.org/10.1002/rth2.12098
10. Daconta, M. C., Obrst, L. J., & Smith, K. T. (2003). **The Semantic Web: a guide to the future of XML, Web services, and knowledge management**. John Wiley & Sons
11. Fensel, D., Ding, Y., Omelayenko, B., Schulten, E., Botquin, G., Brown, M., & Flett, A. (2001). **Product data integration in B2B e-commerce. IEEE Intelligent Systems**, 16(4), 54-59. https://doi.org/10.1109/5254.941358
12. Giri, K. (2011). **Role of ontology in semantic web**. DESIDOC Journal of Library & Information Technology, 31(2). https://doi.org/10.14429/djlit.31.2.863
13. Gormley, C., & Tong, Z. (2015). Elasticsearch: **The definitive guide: A distributed real-time search and analytics engine**. " O'Reilly Media, Inc.".
14. Grainger, T., Potter, T. and Seeley, Y., 2014. **Solr in action**. Cherry Hill: Manning.
15. Hepp, M. (2005). **eClassOWL: A fully-fledged products and services ontology in OWL.** Poster Proceedings of ISWC2005. Galway.
16. Hepp, M. (2008, September). **Goodrelations: An ontology for describing products and services offers on the web. In International Conference on Knowledge Engineering and Knowledge Management** (pp. 329-346). Springer, Berlin, Heidelberg.
17. Hepp, M., & Radinger, A. (2010). **eClassOWL-The Web Ontology for Products and Services. OWL representation of the eCl@ss classification standard**. URL: http://www. heppnetz. de/projects/eclassowl (09.07. 2012).
18. Madhu, G., Govardhan, D. A., & Rajinikanth, D. T. (2011). **Intelligent semantic web search engines: a brief survey**. arXiv preprint arXiv:1102.0831. https://doi.org/10.5121/ijwest.2011.2103
19. McGuinness, D. L., & Van Harmelen, F. (2004). **OWL web ontology language overview**. W3C recommendation, 10(10), 2004.
20. MINCH, A. (2014). **ENGINEERING SEMANTIC WEB FOR E-COMMERCE BUSINESS INTELLIGENCE: A BILINGUAL EEPS ONTOLOGY MODEL** (Doctoral dissertation, ARBA MINCH UNIVERSITY).
21. Pauwels, P., Zhang, S., & Lee, Y. C. (2017). **Semantic web technologies in AEC industry: A literature overview**. Automation in Construction, 73, 145-165.
22. Peng, P., Zou, L., Özsu, M. T., Chen, L., & Zhao, D. (2016). **Processing SPARQL queries over distributed RDF graphs.** The VLDB Journal—The International Journal on Very Large Data Bases, 25(2), 243-268.
23. Pérez, J., Arenas, M., & Gutierrez, C. (2006, November). **Semantics and Complexity of SPARQL.** In

International semantic web conference (pp. 30-43). Springer, Berlin, Heidelberg https://doi.org/10.1007/11926078_3

24. Rebstadt, J., Brinkschulte, L., Enders, A., & Mertens, R. (2016, September). **A Visual Language for OWL Lite Editing.** In SEMANTiCS (Posters, Demos, SuCCESS).
25. Rudman, R., & Bruwer, R. (2016). **Defining Web 3.0: opportunities and challenges**. The Electronic Library, 34(1), 132-154.
26. Stadler, C., Arndt, N., Martin, M., & Lehmann, J. (2015). **RDF Editing on the Web. In SEMANTiCS** (Posters & Demos) (pp. 96-99)
27. VijayaLakshmi, B., GauthamiLatha, A., Srinivas, D. Y., & Rajesh, K. (2011). **Perspectives of Semantic Web in e-commerce.** International Journal of Computer Applications, 25(10), 52-56. https://doi.org/10.5120/3172-4166
28. Wolff, B. G., Fletcher, G. H., & Lu, J. J. (2015, March). **An Extensible Framework for Query Optimization on TripleT-based RDF Stores**. In EDBT/ICDT Workshops (pp. 190-196).
29. World Wide Web Consortium. (2013). **RDFa 1.1 primer: rich structured data markup for web documents**.
30. Sander, M., Waltinger, U., Roshchin, M., & Runkler, T. (2014, September). **Ontology-based translation of natural language queries to SPARQL**. In 2014 AAAI Fall Symposium Series.
31. Elmadany, Hassan A. & Aref, Mostafa & Alfonse, Marco. (2017). (2017) **A Semantic Framework for Summarizing XML Documents**. International Journal of Advanced Trends in Computer Science and Engineering. 6.
32. Banane, Mouad. (2019). **Querying massive RDF data using Spark**. International Journal of Advanced Trends in Computer Science and Engineering. 1481-1486. 10.30534/ijatcse/2019/68842019.