Knowledge Harvesting for Business Intelligence

Nesrine Ben Mustapha, Marie-Aude Aufaure

1

Knowledge Harvesting For Business Intelligence

Nesrine Ben Mustapha and Marie-Aude Aufaure

Ecole Centrale Paris, MAS Laboratory {Nesrine.Ben-Mustapha, Marie-Aude.Aufaure}@ecp.fr

Summary. *With the growth rate of information volume, information access and knowledge management in enterprises has become challenging. This paper aims at describing the importance of semantic technologies (ontologies) and knowledge extraction techniques for knowledge management, search and capture in e-business processes. We will present the state of the art of ontology learning approaches from textual data and web environment and their integration in enterprise systems to perform personalized and incremental knowledge harvesting.*

Key words: ontology, semantics, ontology learning, knowledge engineering, business intelligence

1.1 Introduction

Over the past few years and with the continuous and rapid growth of information volume, information access and knowledge management, in the enterprises and on the web have become challenging. Besides, the growth rate of data repositories has been accelerated to the point that traditional Knowledge Management System (KMS) no longer provides the necessary power to organize and search knowledge in an efficient manner.

Business Intelligence(BI) solutions offer the means to transform data into information and capture knowledge through analytical tools for the purpose of enhancing decision making. Analytical tools are quite dependent on knowledge representation, capture and search. Despite the progress on these analytic tools, there are many challenges related to knowledge management that should be tackled. We argue that these issues are due to the lack of integrating the engineering of business' semantics in the foundation of BI solutions. Therefore, the improvements on knowledge engineering, search and capture offer new opportunity for building enhanced semantic BI solutions.

In fact, if semantic content of resources is described by keywords or natural language or metadata based on predefined features, it is hard to manage it because of its diversity and the need for scalability. Thus, adding a semantic layer

that provides common vocabulary to represent semantic contents of resources contributes to enhance knowledge management, update, sharing and retrieval among different domains. In the emerging semantic web, Information search, interpretation and aggregation can be addressed by ontology-based semantic mark-up.

In this paper, we outline the need of *semantic technologies* for business intelligence in Section 1.2. After studying some motivating use case, we will detail the evolution of correlated dimensions that have been od interests by academic research groups and which include *search technologies* presented in Section 1.4, *Ontology learning approaches* from textual data and web environment and the machine learning techniques detailed in Section 1.5. In Section 1.6, we emphasize the importance of ontology technology and search solution capabilities for semantic revolution of the BI. Finally, we conclude with a brief synthesis.

1.2 Need of Semantic Technologies for Business Intelligence

We remind that the main goal of this lecture resides is explicitly transfer semantic technologies from the field of academia to industry. In this section, we will outline, the central role of knowledge in Business enterprises and the motivating scenarios of integrating semantic technologies in BI processes.

1.2.1 Knowledge Groups in BI Environment

In business enterprises, we can distinguish five main knowledge groups:

- knowledge workers;
- knowledge exploiters;
- knowledge learners;
- knowledge explorers;
- knowledge innovators.

Knowledge workers have an important and internal role in business enterprises. They should have great communication, learning, acting and resolving skills in order to empower the strategy of planning, sharing and collaboration of the enterprise. This group focuses mainly on internal business process and knowledge.

The main priority of knowledge exploiters resides on external knowledge learning than internal one, since they focus on competition knowledge and the development of new product. In order to achieve this purpose, this group should search daily for the external knowledge about competitors strategies, client satisfaction, etc.

The knowledge learners group aims to learn knowledge in certain areas and is not able to integrate different streams of knowledge. So, it is considered slow in learning new knowledge.

The knowledge explorers group has a central role in business enterprises, since it should maintain a good balance between internal knowledge and external knowledge group.

Knowledge innovators are qualified by "aggressive learners" as they try to combine external and external knowledge learning in order to research and disseminate findings from enterprises resources.

In the next subsection, basing on real use cases that have been studied by knowledge-Web network [1], the need of semantic technologies in business enterprise is explained.

1.2.2 Motivating Use Cases: Need of Semantic Technologies

We distinguish three industrial fields for which the need of semantic technologies were discussed in a survey [80], as follows:

- Service industry;
- Media field;
- Health services.

In the following subsections, we will detail problems faced in mentioned uses cases and we will outline possible semantic architectures that can be set up.

Semantic Technologies for Service Industry

We have considered in service industry two main use cases that have been studied in [80]:

- Recruitment use case;
- B2C market place for tourism.

The recruitment service of employees on the web is an emerging and important practice in many business fields. While classic appliance ways remain available (newspaper advertisements, human resource department, etc), the internet has evolved into an open recruitment space. In Germany, 50% of recruitment are expected to be made through online job posting.

Current recruitment systems (such as Monster [2], experteer [3] and jobpilot [4], etc.) provide abilities to publish and search vacancies in addition to posting applicant CVs. The search process on these systems is based on predefined

[1] http://www.knowledgenetworks.com/
[2] http://www2.monster.fr/
[3] http://www.experteer.fr/
[4] http://www.fr.jobpilot.ch/

criteria (skills, job location, domain, etc.). So, the new challenge of these systems is to improve facilities of efficiently filling open job vacancies with the suitable candidates.

In other words, an automatic matching between job offers and job seekers information can be a good solution to overcome this challenge. In this kind of solution, an exact matching of words will not bring remarkable advantages. Integrating *semantic representation* of filled data from job seeker and job provider and *semantic search solution* can cover open issues by purposing the semantic web as technological solution.

According to this use case, we can imagine a possible architecture that integrates an ontology-based support for (Fig 1.1):

- expressing relationships between job characteristics and candidate qualifications (knowledge base);
- semantic querying of job offers or suitable candidates;
- learning knowledge from the web (metadata crawler);
- semantic matching.

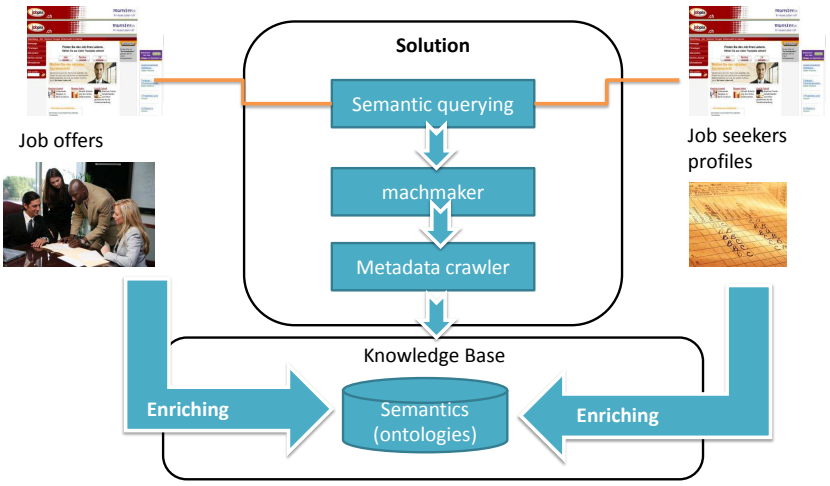

Fig. 1.1. Semantic solution for recruitment use case

Besides, in tourism domain which is a network business, business process relies on a number of stakeholders to develop and deliver products and services. Networks associated with Web 1.0 and Web 2.0 are confronted by serious challenges since limited interpretability is provided. In fact, dedicated sites for regional tourism have substitute the knowledge workers by content management capabilities. The choice of research criteria on B2C marketplace for tourism is limited to predefined suggestions. They are also based on prepackaged offers.

Meantime, travel consumers are now asking for more complex packages involving many itineraries and engaging extensively through online searches in order to meet their information needs. With the current systems based on web 1.0, two main problems are inhibiting the achievement of these new challenges, which are:

- static management of web site content;
- static search process with no personalization.

Therefore, to provide a personalized service including an integrated cost of the involved services (hotels, restaurants, train, plane, geo-localization), new requirements should focus on providing:

- a web content aggregation platform;
- Dynamic exploitation of content, service providers and personalized data;
- Geo-localization;
- on-line personalized tourism packages.

Semantic Technologies for Media Field

With the continuous growing of multimedia databases, it becomes crucial to manage and maintain this huge data set. Classic solutions include faster hardware and sophisticated algorithms. Rather a deeper understanding of media data is needed to perform the multimedia content organization, reuse and distribution.

Since the semantic web is based on machine-processable knowledge representation, there is growing interest on:

- semantic annotation of multimedia content;
- knowledge extraction from media data;
- semantic search, navigation and reasoning capabilities.

One one hand, some projects such as the aceMedia project [5] focuses on discovering and exploiting knowledge inherent to the content in order to make searched content more relevant to the user.

On other hand, others project of news aggregation such as Neofonie [6] are focusing on integrating semantic technologies in order to perform automatic integration and processing of news sources through a thematic clustering techniques.

Semantic Technologies for Health Services

In the context of health care, lack on data management capabilities might lead to a dramatic restructuring of the service and cost model. Doctors may

[5] www.aceMedia.org
[6] http://www.newsexpress.de

ask for remote diagnosis in order to access to the accumulated knowledge of every known example of your symptoms, and your entire medical history from the time you were born.

With the digitalization of medical and health information, the doctor will be able to access to the records of all your prior treatments, including heterogenous data: images, test results, drug reactions and practitioner opinions. Therefore, he can act quickly in order to determine the suitable medications.

However, in the most occurring cases, health care organizations such as hospitals may have several information completely dispersed and not easily reused for other organizations. The main challenges are that:

- Data should be integrated independently from the data type (structured or unstructured source);
- Large health insurance companies use a cognos data warehousing solution to administrate its data;
- Business data are stored in various machines and don't share the same data formats.

Consequently, only manual search over data sources is available. For this reason, introducing common terminology for health care data and solving problems of updating data are requested. A semantic solution will involve three main actors:

- Data architect;
- Knowledge engineer;
- knowledge explorers.

The idea is to build ontologies that can be used for integrating heterogenous data marts into a single consolidate datawarehouse, as shown by figure 1.2.

The overall challenge of these use cases resides in identifying where knowledge is located, how to leverage it for business purpose by harvesting knowledge from enterprise sources and from competitor events and how to manage it in an optimal way. Therefore, semantic knowledge representation is the key for the development of present intelligent systems. In the next subsection, correlated dimensions to the evolution of semantic technology are discussed since it is important to understand how these technologies have appeared in order to be able to choose the adequate technique for a given challenge faced by an industrial organization.

1.2.3 Correlated Dimensions Affecting Semantic Technologies

As stated in Figure 1.3, the main dimensions that affect the evolution of Business Intelligence Solutions are mainly:

- Semantic technologies;
- Structure of the web;

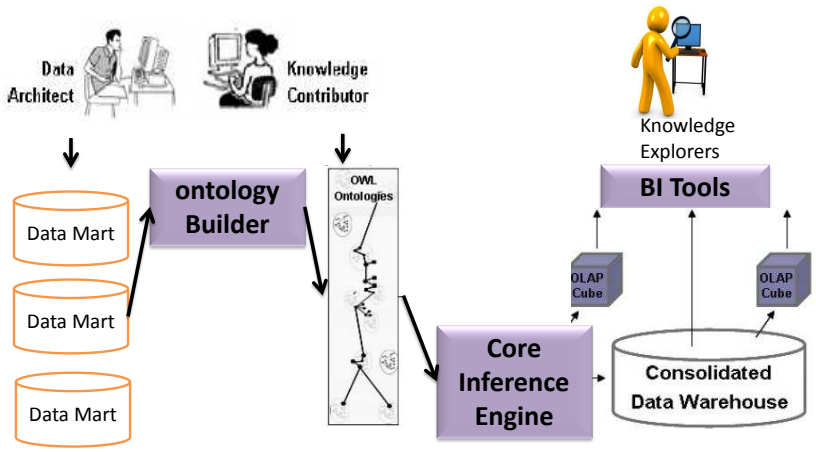

Fig. 1.2. Semantic solution for health use case

- Search methods;
- Knowledge engineering approaches.

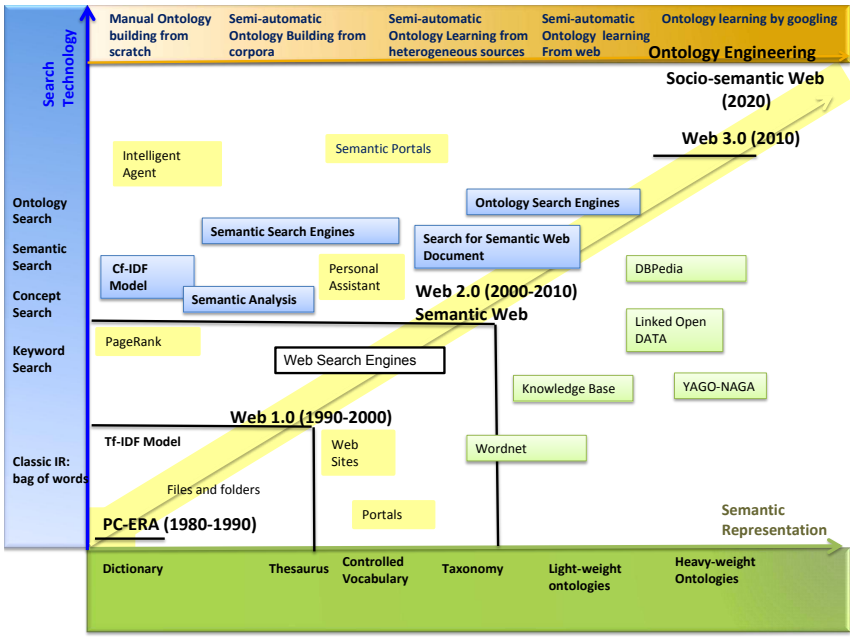

Fig. 1.3. Dimensions affecting Business Intelligence evolution

These dimensions are quite correlated to each other and pave the way towards Business Intelligence 2.0.

Search technologies have evolved from simple keyword searching to relevance ranking (like Google). Text mining, also called text analysis, analyzes unstructured content to better determine the context and meaning of the information in relation to the search terms while relevancy ranking looks at popularity and linkages to other documents. For instance, IBM can search unstructured data sources, use text mining to extract relevant information, and load appropriate contents back into data warehouses.

With the emergence of the semantic web, *Knowledge Representation* methods have evolved from dealing with controlled vocabularies, dictionaries to managing domain ontologies. Domain ontologies have become essential for managing increasing resources (contents) and promoting their efficient use for many software areas such as Bioinformatics [1], educational technology systems [2], E-learning [3], ontologies for commerce and production organization (TOVE [4] and Enterprise [5]), museums and cultural heritage and physical systems, etc.

1.3 Evolution of Semantics: From Dictionaries To Ontologies

As shown in Figure 1.3, The Knowledge Representation area has known several levels of formalization before the incoming semantic web. This is having continuously direct impact on the progress of Information Retrieval and Knowledge Engineering areas.

1.3.1 Levels of Knowledge Specification

Several levels of knowledge formalization can be identified, from controlled vocabularies to heavy ontologies [6], as represented in Figure 1.4:

- **Controlled vocabularies**: is a set of terms defined by domain experts. The meaning of words is not necessarily defined and there is no logical organization between the terms. The vocabulary can be used in order to label documents contents. Catalogs are examples of controlled vocabularies.
- Another potential specification is a **glossary** (a list of terms and meanings): the meanings are specified by natural language statements. This provides more information since humans can read the natural language statements. Typically interpretations are not unambiguous and thus these specifications are not adequate for computer agents.
- A **thesaurus**: provides additional semantic with a limited number of relations such as synonymy (preferred term, term to use instead), related terms (a term more specific, more generic term, term related). These relationships may be interpreted unambiguously by computer agents, but no

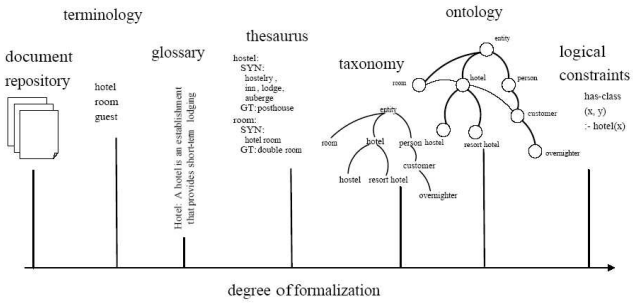

Fig. 1.4. Evolution of knowledge formalization

explicit relationships are specified, although with narrower and broader term specifications, a hierarchy can be deduced.

- **Informal Taxonomy**: provides explicit organizing categories from general concepts to specific ones. They have appeared on the web such as the hierarchy proposed by Yahoo for the categorization of domain topics. However, these hierarchies are not formal because the hierarchy of categories does not meet the strict notion of subsumption .
- Beyond informal "is-a" taxonomy, we move to **formal "is-a" hierarchies**. These include strict subclass relationships. In these systems, if A is a superclass of B, then if an object is a subclass of B, it is necessarily a subclass of A too.

With the emergence of the semantic web, an important trend of ontology-based application has appeared. Definitions and typology of ontological knowledge are detailed in the following subsections.

Ontology Structure

In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or concepts), attributes (or properties), and relationships (or relations among class members). The definitions of these representational primitives include information about their meaning and constraints on their logically consistent application.

The definition adopted by the community of knowledge engineering is the one proposed by Gruber who defines an ontology as "*an explicit specification, a formal shared conceptualization*" [8]. The *conceptualization* is the formulation of knowledge about a world in term of entities (objects, relations between these objects and the constraints of restrictions on these relations). The *specification* is the representation of the *conceptualization* in a concrete form using a knowledge representation language. In the literature, the definition and the

structure of ontologies depend on the type of knowledge, the domain of knowledge and especially the usage of the ontology. In general, the structure of the ontology is composed of:

- a set of *concepts*;
- a set of *taxonomic relationships* between concepts;
- a set of *non-taxonomic relationships*;
- a set of *instances* assigned to each concept or relation;
- a set of *rules or axioms* that represent inferred knowledge of the domain.

These elements are described in the following subsections.

Concept

A Concept is defined as a class of objects related to a well-defined knowledge area such as the concept of "*human being*", *tree*, *home*, etc. It is characterized by its meaning referred by "*Concept intension*" and by its group of entities referred by "*Concept extensions*".

All the objects related to a concept build the set of its instances. For example, the term "*automobile*" refers both to the concept "*car*" as an object of type "*car*" and all objects of that type. A term can refer to several concepts related to several domains. An example of the term "*accommodation*" which refers to the concept of hosting web sites in the topic of "*creation of web pages*" and also the concept of *hotel accommodation* in the field of "*tourism*". Similarly, the same concept can be referenced by multiple terms. This is the case of "*morning star*" and "*evening star*", which both refer to Venus planet [7].

Relations

Within an ontology, we distinguish two main categories of relations: *taxonomic relations* and *non-taxonomic relations*. *Taxonomic* relations organize concepts in a tree structure and include two types of taxonomic relationships:

- relations of *hyponymy* or specialization generally known as "*is a* relation". For example, an enzyme is a type of protein, which is a kind of macromolecule;
- partitive relations or *meronymy* that describe concepts which are part of other concepts.

On the other hand, *non-taxonomic relations* include:

- *locative relation* that describes the location of a concept. Example: "bed *is_located_in* bedroom";
- *associative relations* that correspond to properties between concepts. Logical properties are associated with these relations such as transitivity and symmetry.

Defining the ontology only by concepts and relationships is not enough to encapsulate knowledge, since according to S. Staab and A. Maedche, the axioms and rules are a fundamental component of any ontology [9].

Axioms and Rules

Ontology *axioms* provide semantics by allowing ontology to infer additional statement. Ontological knowledge can be considered as facts, rules, or constraints. A fact is a true statement, not implicative. For example, the axiom "*the company E has 200 employees*" is a true statement. They are useful for defining the meaning of the components, setting restrictions on attribute values, specifying the arguments of a relationship and verifying the validity of specified knowledge.

In recent projects, *axioms* have been extended with *rules* in order to infer additional knowledge. For instance, the following rules "*if a company sells X products A and the price of each product is C then Sales revenue is C * X euros*" is used to calculate the revenue of daily sales.

Several languages have been developed to specify ontology *rules*. Among these languages, we cite mainly RuleML [10] and SWRL [11].

Figure 1.5 illustrates the formalization of the following rule specifying family relatedness using SWRL:

Rule: "hasParent (?x1,?x2) $\wedge$ hasBrother (?x2,?x3) $\longrightarrow$ hasUncle (?x1,?x3)"

Ontology Types

The literature defines four typologies according to the following dimensions:

- formalization degree of the ontology language (formal, informal, semi-formal);
- granularity degree of the ontology structure (light-weight ontology and heavy-weight ontology);
- level of completeness;
- type of domain knowledge.

Typology according to the type of domain knowledge was the most discussed one by several works. We distinguish the following ontology types:

- **Representational ontology**: includes primitives involved in the formalization of knowledge representation. We cite for example, the Frame ontology where primitive representation of language-based Frame are classes, instances, aspects, properties, relationships and restrictions;
- **Top-level ontology (also known as upper ontologies)** specifies concepts that are universal, generic, abstract and philosophical. It provides a set of general concepts from which a domain ontology can be constructed.

```
<ruleml:imp>
 <ruleml:_rlab ruleml:href="#example1"/>
 <ruleml:_body>
  <swrlx:individualPropertyAtom swrlx:property="hasParent">
   <ruleml:var>x1</ruleml:var>
   <ruleml:var>x2</ruleml:var>
  </swrlx:individualPropertyAtom>
  <swrlx:individualPropertyAtom swrlx:property=" has_brother">
   <ruleml:var>x2</ruleml:var>
   <ruleml:var>x3</ruleml:var>
  </swrlx:individualPropertyAtom>
 </ruleml:_body>
 <ruleml:_head>
  <swrlx:individualPropertyAtom swrlx:property="has_uncle">
   <ruleml:var>x1</ruleml:var>
   <ruleml:var>x3</ruleml:var>
  </swrlx:individualPropertyAtom>
 </ruleml:_head>
</ruleml:imp>
```

Fig. 1.5. Rule specification

Examples of existing upper ontologies include SUMO (Suggested Upper Merged Ontology)[7], the CYC ontology[8], and SUO 4D ontology[9];

- **Lexical ontology**: is an ontology describing linguistic knowledge, which models the meaning of words by using ontological structure. Examples of lexical ontologies are WordNet[12] and HowNet [13];
- **Domain Ontology** is tied to a specific domain which can be extended from upper ontology. Examples of domain ontologies include MENELAS in the medical field, ENGMATH for mathematics and TOVE in the field of enterprise management;
- **Ontology of tasks** [14]: used to conceptualize specific tasks such as diagnostic tasks, planning tasks, design tasks, configuration and solving problems tasks;
- **Application ontology** defines the structure of knowledge necessary to accomplish a particular task.

[7] http://suo.ieee.org/SUO/SUMO/index.html
[8] http://ww.cyc.com
[9] http://suo.ieee.org/SUO/SUO-4D/index.html

1.4 New Trends of Search Paradigm

Information Retrieval (IR) research has moved from syntactic IR to semantic IR. In syntactic IR, terms are represented as sequences of characters and IR process is based on computation of string similarity. The progress made by knowledge representation and the semantic web languages areas has contributed to the development of semantic IR systems. Instead, terms are represented as concepts and IR is performed through the computation of semantic relatedness between concepts.

1.4.1 Semantic Web Search in Web 2.0

Several classifications of search engines for the semantic web have been proposed in the literature. Indeed, in [15], the authors distinguish:

- *Document oriented* search engines;
- *Entity oriented* search engines;
- *Multimedia search* engines;
- *Relation Oriented* Search;
- Search Engine based on *semantic analysis*.

Based on semantic search survey presented in [16], we distinguish two types of search engines for the semantic web (Figure 1.6):

- ontology search engines;
- semantic search engines: the use of contextual information (represented by domain ontologies and metadata) is one of the key aspects for these engines.

The two classifications are quite complementary. Ontology search includes entity oriented search and relation oriented search. Semantic search includes the other categories.

Ontology Search Engines

We distinguish two categories of ontology search engines:

The first type corresponds to engines providing specific types of files (such as RSS, RDF, Owl) and enabling to search only by the name of files or by using some options like *filtetype*. For example, in [17], *OntoSearch* engine transmits the user request to Google to search for a specific type of file and uses a visualization tool that allows the user to run her research and to show results. In [18] a technique called *Swangling* is used for this purpose. This technique offers the translation of RDF triples into strings to be transmitted to a traditional search engine. The main problem of these systems is that a lot of available semantic web documents (ontologies) are ignored. In fact, it is not obvious to collect all ontologies in the web just by using filetype command within existing commercial search engines.

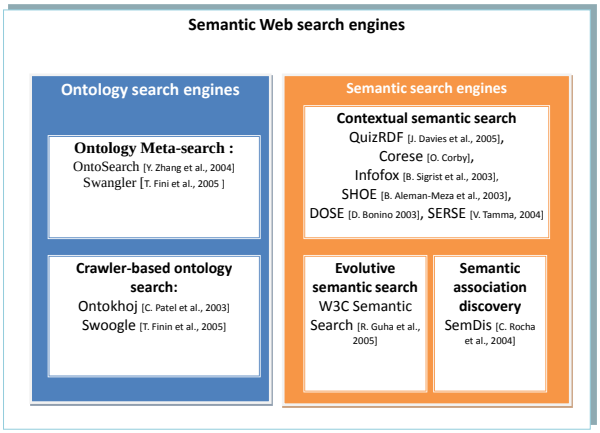

Fig. 1.6. Semantic web search Engine categories

The second type refers to crawler-based ontology search engines. The idea of these system is to build a specific crawler which is used to find Semantic Web Documents (SWD) on the web, index them and acquire some metadata about them. Ontokhoj [19] and Swoogle [18] are two crawler-based ontology search engines. By using these engines users can search for special class, property and entities.

Semantic IR Engines

The following groups can be distinguished:

- Contextual search engines;
- Evolutionary Search engines;
- Semantic association discovery engines.

Contextual Search Engines

The ultimate goal of these engines is to increase the performance of traditional search engines (especially in regard to measures of precision and recall). The use of contextual information (represented by a domain ontology and metadata) is one of the main aspects. Usually after a traditional search process, matching RDF graphs is used to obtain better results.

We distinguish seven major components: crawler, documents annotator, indexer, query formulation module, query annotation module, search module and display module. Various approaches and solutions for each of these subproblems have been proposed [3]. It should be stressed that a very limited number of engines include all the components listed above. The quality of the results depends heavily on ontologies used. The main problem of these

search engines relies with the fact that their use is limited to specific domains (represented by domain ontologies).

The best known examples are: OWLIR [20], QuizRDF [21], InWiss [22], Corese [23], SHOE [24], DOSE[25], OntoWeb [26], SERSE [27].

Evolutionary search engines

In the second group employing semantic web techniques, the objective is to accumulate information on a subject that we seek. This type of search engines is a response to a well-known problem: the automatic collection of information on a domain. The originality of these engines is the use of external metadata (eg. CDnow, Amazon, IMDB). They are usually employed a conventional search engine and provide regular information in addition to the original results: W3C Semantic Search [28] and ABC [29].

Semantic association discovery

Search engines in the third group try to find semantic relationships between two or more terms: the aim is to find various semantic relations between the terms of entries (usually two) and then rank the results based on semantic distances. Compared to other categories, the engines dedicated for discovering semantic associations are linked to higher layers of semantic web architecture (logic and trust). SemDis [28] is an example of this group.

1.5 Progress in Ontology Engineering Research

The methodologies proposing manual ontology building, also known as "*from scratch*" were among the first works done in the field of ontology engineering. It consists in conceiving a process of ontology building in the absence of *a priori* knowledge (hence the meaning of the English term "*from scratch*"). Several authors have proposed many approaches based on learning techniques in order to improve the automation of this process.

The notion of learning reinforces the idea of ontology construction on the basis of *a priori* knowledge. This allows the automation of the ontology enrichment by using learning techniques. According to Maedche and Staab, there are as many ontology learning approaches as types of data sources [30]. We distinguish ontology learning approaches from texts, from dictionaries [31], from knowledge bases [32], from semi-structured schema [33] and relational data [34]. In this section, we are interested mainly in approaches related to ontology learning from web (including texts).

1.5.1 Ontology Learning Approaches

Ontology learning (OL) is defined as an approach of ontology building from knowledge sources using a set of machine learning techniques and knowledge

acquisition methods. OL from texts is a specific case of OL from web and has been widely used in the community of knowledge engineering since texts are semantically richer than other data sources. These approaches are generally based on the use of textual corpora. This one should be a representative of the domain for which we are trying to build ontology. By applying a set of text mining techniques, a granular ontology is enriched with discovered concepts and relations from textual data. In such approach, human intervention is required to validate the relevance of learned concepts and relations.

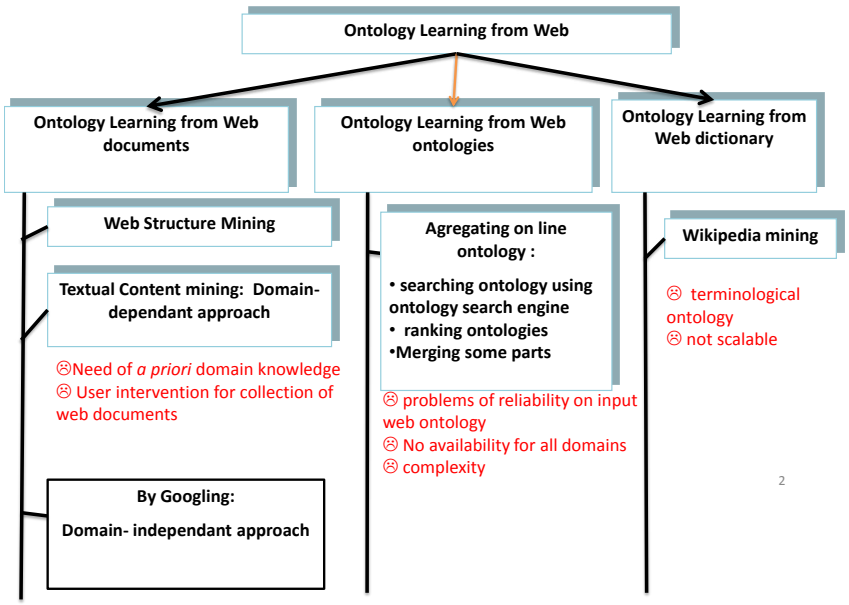

Fig. 1.7. Ontology Learning Approaches from web

In the last decade, with the enormous growth of online information, the web has become as an important data source for knowledge acquisition due to its huge size and heterogeneity. This led to mainly five categories of OL approaches:

- Ontology learning based on web content mining (texts);
- Ontology learning based on web structure mining;
- Ontology learning from web dictionary;
- Ontology learning from web ontologies;
- Ontology learning by googling.

Ontology Learning Based on Web Content Mining (Texts)

OL approaches from texts have widely interested the ontology engineering community. These approaches are based on machine learning techniques applied to texts. Ontology learning process from texts consists generally in enriching a small ontology called "minimal" or "granular" ontology with new discovered concepts and relationships from input texts (corpora or web content document). This is in particular the work of: [35] [36] [39] [40] [41] [42] [43] [44] [45] [46].

By using a set of text mining techniques, knowledge contained in texts is projected to the ontology by extracting concepts and relations. We distinguish mainly five categories of techniques:

- Linguistic techniques [39] and lexico-syntactic patterns [38];
- Clustering techniques and / or classification techniques [35] [36] [67];
- Statistical techniques [47] [49] ;
- Association rule-based techniques [9];
- Hybrid ones.

Sekiuchi98

Ontology Learning Based on Web Structure Mining

Furthermore, others researchers were interested to study the structure of a growing number of web pages. The underlying assumption behind **web structure mining-based** is that the noun phrases appearing in the headings of a document as well as the document's hierarchical structure [50] can be used to discover taxonomic relations between concepts.

Several systems supporting this approach analyze input documents' heading structure, extract concepts from headings and builds a taxonomical ontology. [51] defines an approach for an automated migration of data-intensive web sites into the semantic web. It is based on the extraction of light ontologies from structured resources such as XML Schema or relational database schemata and consists in building light ontologies from conceptual database schemas using a mapping process. This process provides the conceptual metadata of annotations that are automatically created from the database instances.

Besides, in [52] an approach called "Tango" uses the analysis of tables in web pages for the generation of ontologies. In these works, the main difficulty resides on the interpretation of the HTML structure that cannot reflect the semantics of documents. Human intervention is still necessary to validate the resulted ontologies.

Ontology Learning From Web Ontology

With the development of standards and tools supporting the semantic web vision, harvesting ontological files on the web has been the first step towards

achieving true ontology reuse for ontology learning. The idea about online **ontology building from web ontology** has widely been explored by several works [51] [53] [54]. However, the objective was mainly to enable users to reuse or import whole ontologies or ontology modules. They provided no support for ranking available ontologies, or for extracting and merging the ontology parts of interest, or event for evaluating the resulting ontology. In [53], a framework for integrating multiple ontologies from structured documents into a common ontology is used. A universal similarity paradigm reflecting the implicit cohesion among the ontologies is presented. Ontology alignment and construction methods are applied.

Other approaches use ontology search engines or ontology meta-search engines to build ontologies by aggregating many searched domain ontologies. There is an increasing number of online libraries for searching and downloading ontologies. Examples of such libraries are Ontolingua, Protege, and DAML. Few search engines have recently appeared that allow keyword-based search for online ontologies, such as Swoogle and OntoSearch.

In [54], the proposed approach consists in searching online ontologies for certain concepts, ranking the retrieved ontologies according to some criteria, then extracting the relevant parts of the top ranked ontologies, and merging those parts to acquire the richest domain representation as possible.

We don't deny that these approaches can easily lead to obtain many domain ontologies but some problems still remain. In fact, we still worry about many issues:

- Existing web ontology are not sufficiently consistent to be used;
- the availability of ontologies to be reused in terms of number and domain variety;
- the quality of output ontology depends on the quality of input ontologies;
- the use of ontology searching, ontology ranking, ontology mapping, ontology merging, and ontology segmentation methods makes this approach more complex.

Ontology Building From Web Dictionary "*Wikipedia mining*" is a research area recently addressed. In [55], a construction method based on Wikipedia mining is proposed. By analyzing 1.7 million concepts on Wikipedia, a very large scale ontology (called "YAGO") which has more than 78 million associations was built. To avoid natural language processing (NLP) problems, structure mining is applied to web-based dictionaries [55].

Other hybrid approaches In [30], an approach combining heterogeneous sources of information and various processing techniques associated with each type of data source was proposed in order to improve the identification of potential useful knowledge. First, it extracts the core vocabulary to the domain using a parsing process. The underlying idea of the method is that the combination of all these additional sources of evidence improves the accuracy of the OL process. Thus, the extracted terms are analyzed at five different levels: chunk, statistical, syntactical, visual and semantic level. The experimental

results obtained by processing a set of HTML documents belonging to two domains, *Universities* and *Economics*, have shown the potential benefit of its use to learn or enrich ontologies following an unsupervised learning approach.

1.5.2 Generic Ontology Learning Process From Texts

The extraction process starts from the raw text data (text document in natural language) to obtain the final ontology knowledge representation. It includes the following steps [56]:

- Term extraction;
- Synonym extraction;
- Concept discovery;
- Taxonomic relation learning;
- Non-taxonomic relation learning.

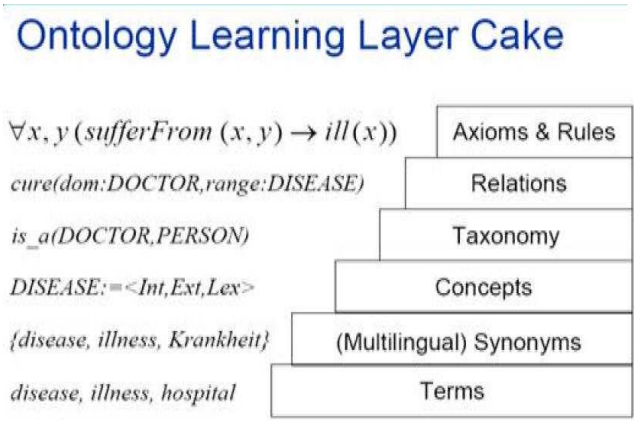

Fig. 1.8. Ontology Learning Process [56]

Term Extraction

A term is a semantic unit and can be simple or complex. The terms are extracted using several techniques including statistical analysis [57], use of patterns (regular expressions), linguistic analysis (identification of nominal and prepositional groups), word disambiguation [58] and interpretation of compound phrases (as in [59] using Wordnet).

Linguistic Techniques of Term Extraction

Linguistic analysis of texts also requires the use of a grammar representing the sentence structure. We distinguish two types of grammars that mainly allow to represent the structure of a sentence in a given natural language:

- *Constituency Grammar*: this grammar is the basis of the formal theory of language used in computational linguistics. The analysis using this type of grammar is based on the position of words in the sentence and how they can be grouped.
- *Dependency Grammar*: the analysis using this grammar provide binary grammatical links between words in a sentence. When two words are connected by a dependency relationship, we say that one is the ruler or the head and the other is a dependent. In general, the extracted relations are schematically represented by an arc between the head and the dependent.

Statistical Techniques for Term Selection

Statistical techniques are mainly based on the analysis of word co-occurrences and other parameters such as absolute frequency of a term, frequency of a term on a given field, etc. Under the assumption of Harris [60], these methods determine a score representing the relationship between two terms and retain those with scores greater than or equal to a given threshold. For example, the combination of TF$*$ IDF measure with other methods such as *latent semantic analysis* can be used to select the domain concepts. It should be noted that these methods ignore the statistically insignificant terms.

The measures used for selection of candidate terms according to their occurrences in the corpus are as follows:

- The TF-IDF measure [61];
- The entropy [62];
- The Relevance to the domain (PD) [63];
- The Consensus in domain [63];
- The pointwise mutual information (PMI) measure [64] (formula 1.1).

The information contents of the concept is defined by its occurrences in the corpus as well as concepts that it subsumes. It aims to use the probability of a concept in a corpus of documents (formula 1.2). The information contents of a concept c is calculated as following [47]:

$$PMI(c) = -log(p(c)) \tag{1.1}$$

where:

$$p(c) = \frac{freq(c)}{N} \qquad and \qquad freq(c) = \sum_{n \in word(c)} \tag{1.2}$$

Synonyms Extraction

The second step aims to identify synonyms among the extracted terms, in order to associate several words with the same concept in the same language [74]. The extraction of synonyms is usually done in two ways:

- Using lexical ontologies such as Wordnet [48];
- Classification techniques which are used to group terms occurring in the same context (eg, co-occurrences of terms).

Concept Learning

Extracted terms are useful to represent the concepts of an ontology. Concept can be discovered using two techniques:

- Construction of the Topic Signature;
- Classification of concepts based on contextual properties of words.

Construction of the Topic Signature

This technique defined in [35] aims to overcome two main limits of lexical ontologies like Wordnet which are the lack of updating links between concepts and the proliferation of different meanings for each concept. This approach proceeds as follows. Firstly, information contained in existing ontology like Wordnet (synonyms of the concepts, hyponyms, antonyms, etc) is used to build requests which are used to search the relevant documents relating to one sense of a given term. The documents related to the same sense of this term are grouped together to form a collections. Secondly, the documents in each collection are processed. Words and their frequencies are extracted by using a statistical approach.

Extracted data from one collection is compared to data in other collections corresponding to the other senses of the same term. The words having a distinctive frequency for one of the collections are grouped in a list, which make up for each sense of a term, the contextual signature *(Topic signature*) generally used in the construction of summaries of texts. Thirdly, for a given word, the concepts associated with their sense are hierarchically grouped. With this intention, various signatures are compared to discover shared words and to determine intersection between the signatures. Many measures are used to calculate the semantic distance. The contextual signatures were evaluated by their application in the task of semantic disambiguation of words. These first contain considerably useful information for this task. However, the evaluation of this method by using Wordnet is not sufficient to conclude by its effectiveness in the case of domain ontology construction.

Classification of Concepts Based on Contextual Properties of Words

This technique is based on the principles of the *Distributive Semantics* which admit that "*the meaning of a word is strongly correlated with the contexts in which it appears*". This assumption can be generalized to cover complex expressions instead of words. The contexts can be formalized in the shape of words vectors, as in the case of semantic signature of subject described in [65].

By using the *Topic signatures*, each concept is represented by a set of co-occurring words and their frequencies. Within this framework, several metrics of similarity, such as TF∗IDF ou Chi-Square, can be used to measure the distance between various concepts. An algorithm of downward classification is described in [36] in order to extend from existing ontologies (such as Wordnet) with the new concepts. In fact, the quality of topics signatures construction described in [35] can be improved by taking in account only concepts belonging to contexts of existing ontology concepts (ie. which have syntactic relationships to the concepts in ontology). For example, it is possible to consider only the list of the verbs for which the concepts are subjects or a direct object, or to consider only the adjectives which modify the concept.

Learning Taxonomic Relations

At this step, two categories of machine learning techniques (linguistic and statistical techniques) can be used. Linguistic techniques of taxonomic relations discovery are based on the definition of lexical-syntactic patterns for extracting hyponymy relations [38]. Several statistical techniques are based on the analysis of word distribution in the corpora.

Lexico-Syntactic Patterns Related to Taxonomic Relations

Lexico-syntactic patterns are based on the study of syntactic regularities between two given concepts. Indeed, it aims to schematize the lexical and syntactic context of taxonomic relations between concepts. This mapping is a lexico-syntactic pattern and permits the retrieval of pairs of words which satisfy this relation from the corpus.

Hearst's Patterns is the basis of several approaches. We illustrate the patterns of hyponymy relations identified by Hearst in the English language in Table 1.1.

In [39], the experimental evaluation of a large number of patterns was done using the Cameleon tool. The results obtained showed that the effectiveness of these patterns and their meanings depend on the corpus. Indeed, the syntactic regularities regarding the relations of hyponymy that were defined do not necessary reflect the relevant relationships in the ontology.

NP such as NP, NP, ... and NP	data warehousing technologies *such as* reporting, ad-hoc querying, online analytical processing (OLAP).
Such NP as NP, NP, ... or NP	*such* supervised machine learning *as* data pre-processing *or* feature selection.
NP, NP, ... and other NP	screen real estate to financial charts, indices *and other* news graphics.
NP, especially NP, NP,... and NP	Accounting, *especially* financial accounting gives mainly past information in that the events are recorded.
NP is a NP	SAS OLAP *is* a multidimensional data store engine

Table 1.1. Hearst's Patterns

Statistical Techniques for Learning Taxonomic Relations

Several statistical techniques are described in the literature for extracting taxonomic relationships between terms. They are based on analysis of co-occurrences between words in documents. The co-occurrence corresponds to the simultaneous occurrence of two words in a text (or window of n words). The set of term co-occurrences is represented by a matrix. This is then used for:

- an hierarchical grouping of words, using automatic classification methods;
- a grouping based on probability measures [66].

In this context, it is also possible to apply hierarchical clustering by using the co-occurrence matrix of words extracted from documents. In the case of a hierarchical cluster, initially, each class is composed of a term. In [66], a rule related to taxonomic relation extraction stipulates that if two concepts were referred by terms that appear in the same documents (in fact in 80 % of cases), then these two concepts are hyponyms. In other words, if a concept X subsumes a concept Y and the documents in which X appears are a subset of the documents including the word Y, then X subsumes Y. Other rules can be discovered according to the corpus. The conditional probabilities depend closely on the selected context which can be a sentence, a web page, or a web site.

Extracting Non-Taxonomic Relations

Another step of ontology learning consists in discovering *non-taxonomic relations* between concepts. A non-taxonomic relation can be extracted using two main techniques:

- conceptual clustering using syntactic frames [67];
- statistical techniques.

Learning Syntactic Frames

Conceptual clustering requires a syntactic analysis of the documents from which we estimate being able to build an ontology. Classes are formed starting from the terms appearing after the same verb and the same preposition. An algorithm of conceptual clustering is applied for this purpose. One difficulty relies in labeling the relations after their discovery.

To solve this problem, two clustering algorithms: "*Asium-Best*" and "*Asium-Level*" based on the extraction of the syntactic frames were proposed by ASIUM approach in [67]. These techniques allow the discovery of non-taxonomic relations between two classes of terms. These relations are labeled according to the verb and the preposition concerned with the syntactic frame. A syntactic frame for the verb "to travel" is illustrated as following: <To travel><subject:human> <by:convey>.

Initially, the parser automatically provides noun expressions associated with the verbs and the clauses. For example, starting from the following syntactic frame, classes are created:

- <To travel> (<subject: Jean>) (<in: means of transport>);
- <To travel> (<subject: David>) (<in: train>);
- <To lead> (<subject: Helene>) (<object: means of transport> ;
- <To lead> (<subject: Roland>) (<object: plane>).

The classes are successively aggregated to form new concepts hierarchies. The obtained classes are labeled by an expert to identify the concepts which they represent. The classes make up the groupings of words having the same frame: <verb> <syntactic role —preposition: name>, such as for example " <travelling> <subject: human> <by: convey>.

The couples <syntactic role: name> or <preposition: name> are called "heads words". Similarity measures permit to evaluate the distance between the classes, and thus to gather their dependencies based on the proportion of common "heads words" and their frequency of appearance in the documents. The method was tested on a corpus related to kitchen recipes. When the system is involved to find the couples verb-argument on 30% of the corpus, the hierarchy suggested is valid to 30%.

Statistical Techniques For Learning Non-Taxonomic Relations

The main idea of this technique is to extract noun phrases and proper names that appear frequently. These terms are called the central terms. According to the approach DOODLE II [49], co-occurring terms are proposed to be related in the ontology, and the verbs that occur in the context are proposed to be the labels of the relationship. These terms may be determined from one co-occurrence matrix in a window of n words. The advantage of this approach is that preprocessing of texts is avoided and the combination of association rules with the space of words gives better results than each technique separately used.

Finally, recent approaches which propose the use of search engines to learn ontology are described in the next section.

1.5.3 Ontology Capture by Googling

[68] proposes to construct an ontology by submitting the initial keywords to Google in order to retrieve web pages containing these terms. A study of several available types of search engines on the web has been carried out in order to be used in the learning process (searches web resources and calculates a score based on the number of hits) (figure 1.9).

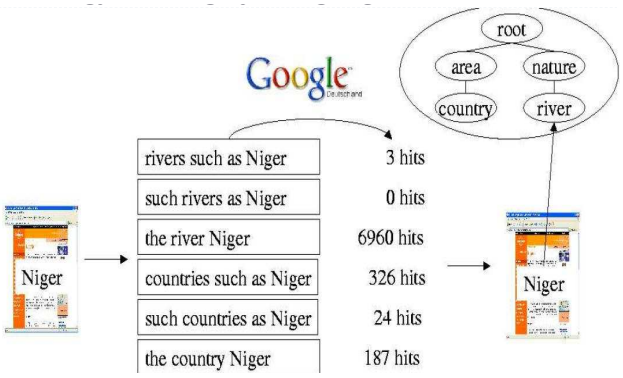

Fig. 1.9. Ontology leaning by googling

The learning process proposed in this approach is based on four steps. The first one is a taxonomic learning step where the user starts to specify a keyword used as a seed for the learning process, using a web search engine. The output of this step is a one-level taxonomy and a set of verbs appearing in the same context as extracted concepts.

Secondly, non-taxonomic learning is carried out. Verb list and keywords are used as a bootstrap for the construction of domain patterns in order to submit reformulated queries to the search engine.

The third step is the recursive learning task where the two previous learning tasks are recursively executed for each discovered concept. Finally, the post-processing step consists in refining and evaluating the obtained ontology. This approach is domain independent and incremental.

These approaches led to identify three main techniques that were adapted to the web:

- statistical techniques based on term-distribution in the web;
- ontology population by Googling;
- label learning for non-taxonomic relations.

Considering the web as a massive source of knowledge, *several statistic approaches* have exploited the number of pages returned by a web search engine to estimate the probabilities of co-occurrence of terms. We consider the following notations:

- $hit(a)$ is used to denote the number of web pages containing the query returned by a search engine;
- $total_{Webs}$ denotes the total number of pages indexed by the search engine.

From a unsupervised point of view, the statistical estimation of the semantic link between concepts, as proposed in [69] [70] typically uses a measure derived from the following co-occurrence function between two terms:

$$C_k(concept, candidat) = \frac{prob(concept \quad AND \quad candidat)_k}{prob(concept) \times prob(candidat)} \tag{1.3}$$

The Symmetric Conditional Probability (SCP)) [69] can be defined as C_2 and the Pointwise Mutual Information (PMI) [70] as $log_2 c_1$.

The probability "$prob(a \quad AND \quad b)$" is computed using the hit number provided by search engines, as stated by the following formula:

$$prob(a, b) = \frac{hit(a \quad AND \quad b)}{total_{Webs}} \tag{1.4}$$

The score derived from this function was defined by Turney as follows:

$$score(concept, candidat) = \frac{prob(conceptANDcandidat)_k}{prob(concept) \times prob(candidat)} \tag{1.5}$$

The measures proposed by Turney were applied and evaluated in [71]. However, since the semantic content and context of words is not taken into account by these measures, limited performance is observed in [72].

Other approaches were interested in *ontology population by googling*. In fact, Gijs Geleijnse and Jan Korst [73] propose the identification of concept instances using the search engine *Google*. Queries are constructed based on lexico-syntactic patterns defied by *Hearst* [38]. A term is accepted when the number of hits (number of results returned by *Google*) exceeds a given threshold.

The same principle was also explored by [74] in order to extract taxonomic relations and attributes of concepts.

Finally, reference work on *learning non-taxonomic relation* from web has been well detailed in [74] and led to the development of the Pankow system. Pankow also relies on the idea that lexico-syntactic patterns described above can be applied not only to a text corpus, but also in the World Wide Web as in [79].

1.6 Ontologies for Business Intelligence

Business intelligence (BI) is defined as the process of searching, gathering, aggregating, and analyzing information for decision making.

Nowadays, Business Intelligence actors intend to bring together researchers in techniques related to conceptual modeling, **ontology engineering**, **knowledge representation**, and **Information Retrieval** for helping business developers, managers, and analysts involved in the development of BI systems to take benefits from heterogeneous data sources (unstructured and structured) and to facilitate information search.

The aim is to perform discussions on integrating ontologies, modeling languages, and search methods for the engineering of BI systems with the purpose of providing more precise information for the end-user, *bridging the gap between the dimensions that affect the evolution of Business Intelligence.*

Besides, semantic technologies advocated by semantic web[77] have been applied for BI in the context of the MUSING Project [10]. The new trend aims to develop a new generation of BI tools and modules based on *semantic-based knowledge and natural language processing (NLP) technology* to make easier gathering, merging, and analyzing information [76].

On the other hand, Ontology-based Information Extraction (OBIE) is a suitable technique for automatically extracting specific fragments from text or other sources to create records in a database or populate knowledge bases. Without an OBIE system, business analysts have to read hundreds of textual reports, web sites, and tabular data to carry out BI activities and feed BI models and tools.

In this paper we stressed the existing relation between Ontology Learning process (OL) and Ontology-based Information Extraction (OBIE) in academic research areas. This relation can be applied to the context of Business Intelligence.

In [78], authors propose a Semantic Business Intelligence (SBI) architecture that incorporates many features that distinguish it from the existing information management solutions and research. Their work aims at enabling the integration of business semantics, heterogeneous data sources, and knowledge engineering tools in order to support a smarter decision making.

Besides, the CUBIST project [11] (Combining and Uniting Business Intelligence and Semantic Technologies) aims to explore standard approaches known from Formal Concept Analysis (FCA) in order to manage the complexity of the visualizations of concepts (for example, by condensing/clustering the resulting concepts, restrict visualizations by means of sub-dividing data, or filtering data in combination on other semantic query forms like faceted search) in the context of BI.

[10] http://www.musing-project.eu
[11] www.cubist-project.eu/

1.7 Conclusion

The improvement on knowledge engineering, capture and search have contributed to tackle knowledge management in the context of BI. These research areas are quite correlated and can affect positively the development of enhanced semantic Business Intelligence Tools. This paper aims to make these correlation more explicit. A state of the art about knowledge representation, recent ontology engineering approaches and semantic search engine are detailed.

On the base of analyzing ontology-based search engines presented in Section 1.4, we have identified the following problems:

- **the scalability of ontologies**: which makes difficult of handling several domain ontologies being used in semantic BI tools.
- **the problem of query reformulation with the use of several domain ontologies**: this is due to the usual mapping problems between query terms, ontological concepts and terms existing in documents. Indeed, identifying the ontological fragment that can be relevant for query reformulation depends strongly on the structure of the ontology. The use of the superclasses of the key concepts or the attributes in the the query reformulation task does not necessarily improve the relevance of search results. For this reason, the context of ontology-based BI applications including data type , ontology usage, users preferences is important to take into account in the ontology building process.

These problems are quire related to the progress made by ontology engineering approches which have been widely described in this paper.

Approaches for building ontologies from online ontologies described in subsection 1.5 are based on the use of ontology search engines, the classification and the aggregation of the resulting ontologies. These approaches can be easily integrated in the semantic BI architecture but several problems inhibit us to continue exploring this idea, including:

- inconsistency of the existing ontologies in online libraries;
- absence of ontologies related to several domain on the web (especially business domain)
- complexity of ontology classification, ontology alignment, merging and segmentation.

In addition, works proposing the construction of ontologies from online dictionaries allowed obtaining very large terminological ontologies (YAGO). These ontologies are quite useful but their update depends on the contents of these dictionaries, and they can be enriched only if these dictionaries undergo the updates.

For these reasons, we can consider the web as a complementary scalable source that is rich of continuously updated texts, and is covering all areas of knowledge. Using Ontology learning techniques based on googling, it will make

it possible to build an integrated BI solution for incremental ontology learning. On the other side, ontology learning techniques can be applied to unstructured content, representing 80% of enterprise data, to build specific knowledge bases and enhance the search and decision processes. We were primarily interested, in this paper, in studying approaches of ontology learning from web content, since the other approaches are based on limited data sources and do not favor the evolution of the ontologies.

References

1. Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R.: The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. 32, 262–266 (2004)
2. Boyce, S., Pahl, C.: The development of subject domain ontologies for educational technology systems. Journal of Educational Technology and Society (ETS) IEEE 10(3). Journal of Educational Technology and Society (ETS) IEEE 10(3), 275–288, (2007)
3. Holohan, E., Melia, M., McMullen, D., Pahl, C.: Adaptive e-Learning Content Generation Based on Semantic Web Technology. In: Workshop on Applications of Semantic Web Technologies for e-Learning, AIED'05. July 18, Amsterdam, The Netherlands (2005)
4. Fox, M. S. and Barbuceanu, M. : An Organisation Ontology for Enterprise Modeling. In: M. Prietula;K. Carley; L. Gasser (Hrsg.): Simulating Organizations: Computational Models of Institutions and Groups. AAAI/MIT Press, Menlo Park CA, pp. 131–152, (1998)
5. Uschold, M., Grninger, M.: ONTOLOGIES: Principles, Methods and 16 applications, Knowledge Engineering Review, Vol. 11, N. 2, 93–13 (1996)
6. Navigli, R., Velardi, P.: From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions. Frontiers in Artificial Intelligence and Applications 167, 71-89 (2008)
7. Furst, F., Leclre, M., and Trichet, F.: Construction d'une ontology oprationnelle : un retour d'exprience. In Hrin, D. and Zighed, D. A., editors, EGC, volume 1 of Extraction des Connaissances et Apprentissage, Hermes Science Publications, pp. 227–232 (2002)
8. Gruber, T.:Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, special issue on Formal Ontology in Conceptual Analysis and Knowledge Representation. Eds, Guarino, N. and Poli , R. (1993)
9. Maedche, A. and Staab, S.: Ontology Learning. In: Staab, S. and Studer, R. (Hrsg.): Handbook on Ontologies. Springer, International Handbooks on Information Systems , S. pp. 173–190 (2004)
10. Harold, B.: Design Rationale of RuleML: A Markup Language for Semantic Web Rules, In: Proceeding of SWWS'01, pp. 381–401 (2001)
11. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: SWRL: A Semantic Web rule language combining OWL and RuleML, W3C Member Submission. (2004)

12. Miller, G.: Wordnet: A lexical database for English. CACM, 38, Vol. 11, pp. 39–41 (1995)
13. Dong, Z. and Dong, Q.: Hownet and the Computation of Meaning. World Scientific (2006)
14. Chandrasekaran, B., Josephson, J. R., and Benjamins, V. R.: Ontology of tasks and methods, Proceedings of the 11th Workshop on Knowledge Acquisition Modeling and Management, Banff, Canada, pp. 20–26 (1998)
15. Wang, W., Payam, M., Barnaghi, A. B.: Search with meanings : An overview of semantic search systems. Journal of Communications of SIWN, Vol 3, 76–82 (2008)
16. Esmaili, K. S. and Abolhassani, H.: A categorization scheme for semantic web search engines. In 4th ACS/IEEE Int. Conf. on Computer Systems and Applications (AICCSA-06),IEEE, pp. 171–178 (2006)
17. Zhang, Y., Vasconcelos, W., and Sleeman, D.: OntoSearch: An ontology search engine. In: Proceedings of the 24th SGAI International Conference on Innovative, Techniques and Applications of Artifial Intelligence, Cambridge, UK, pp. 81–93 (2004)
18. Finin, T. W., Ding, L., Pan, R., Joshi, A., Kolari, P., Java, A., and Peng, Y.: Swoogle : Searching for knowledge on the semantic web. In Veloso, M. M. and Kambhampati, S., editors, AAAI Press, The MIT Press, pp. 1682–1683 (2005)
19. Patel, C., Supekar, K., Lee, Y., Park, E.K.: OntoKhoj: a semantic web portal for ontology searching, ranking and classification. In: WIDM03, pp. 58–61 (2003)
20. Shah, U., Finin, T., Joshi, A., Cost, R.S. and Mayfield, J., Information Retrieval on the Semantic Web. In 10th International Conference on Information and Knowledge Management, (McLean, Virginia, USA, 2002), pp. 461–468 (2002)
21. Davies, J., Weeks, R. and Krohn, U: QuizRDF: Search technology for the Semantic Web. In WWW2002 Workshop on RDF and Semantic Web Applications, Hawaii, pp. 133–143 (2002)
22. Priebe, T.: INWISS - Integrative Enterprise Knowledge Portal., Demonstration at the 3rd International Semantic Web Conference (ISWC 2004), Hiroshima, Japan, pp. 33–36 (2004)
23. Corby, O., Dieng-Kuntz, R., and Faron-Zucker, C. Querying the semantic web with corese search engine. In de M?ntaras, R. L. and Saitta, L., editors, ECAI, pp. 705–709. IOS Press, (2004)
24. Heflin, J. and Hendler, J. Searching the web with shoe. In AAAI-2000 Workshop on AI for Web Search, pp. 35–40 (2000)
25. Bonino, D., Corno, F., and Farinetti, L. DOSE: A distributed open semantic elaboration platform. In ICTAI, pp. 580–588. IEEE Computer Society, (2003)
26. Spyns P., Oberle D., Volz R., Zheng J., Jarrar M., Sure Y., Studer R. and Meersman R., OntoWeb - a Semantic Web Community Portal. In, Karagiannis, D. and Reimer, U.,(eds.), Proceedings of the Fourth International Conference on Practical Aspects of Knowledge Management (PAKM02), LNAI 2569, pp. 189-200, Springer Verlag (2002)
27. Tamma, V., I. Blacoe, B. Lithgow-Smith, and Wooldridge M.:SERSE: Searching for Semantic Web Content, in R. Lopez de Mantaras and L. Saitta (eds.), In:Proceedings of the Sixteenth European Conference on Artificial Intelligence (ECAI-04), pp. 63–67, (2004)

28. Rocha, C., Schwabe, D., and de Aragao, M. P.: An hybrid approach for searching in the semantic web. In Proc. of 13 h Intl. World Wide Web Conf. (WWW 2004), pp. 374-383, (2004)
29. Halaschek-Wiener, C., Aleman-Meza, B., Arpinar, I. B., and Sheth, A. P.: Discovering and Ranking Semantic Associations over a Large RDF Metabase. In 30th International Conference on Very Large Data Bases, (Toronto, Canada, 2004). pp. 1317–1320, Morgan Kaufmann (2004)
30. Maedche, A. and Staab, S.: Ontology Learning for the Semantic Web. IEEE Intelligent Systems, Special Issue on the Semantic Web, Vol 6(2), 72–79, (2001)
31. Jannink J.: Thesaurus Entry Extraction from an On-line Dictionary. Proceedings of Fusion '99, Sunnyvale CA, (1999)
32. Suryanto, H. and Compton, P.: Discovery of Ontologies from Knowledge Bases. Proceedings of the First International Conference on Knowledge Capture, The Association for Computing Machinery, New York, USA, pp. 171–178, (2001)
33. Papatheodrou, C., Vassiliou, A. and Simon, B.: Discovery of Ontologies for Learning Resources Using Word-based Clustering, EDMEDIA 2002, Copyright by AACE, Reprinted, Denver, USA, (2002)
34. Rubin, D.L., Hewett, M., Oliver, D.E., Klein, T.E. and Altman, R.B.: Automatic data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML. In: Proceedings of the Pacific Symposium on Biology, Lihue, HI, pp. 88–99 (2002)
35. Agirre, E., Ansa, O., Hovy, E. and Martinez, D.: Enriching very large ontologies using the WWW, In Proceedings of ECAI Workshop on Ontology Learning (ECAI-00), (2000)
36. Alfonseca, E., Manandhar, S. : An unsupervised method for general named entity recognition and automated concept discovery ". In Proc. First International conference on genaral WordNet, India, (2002)
37. Ben-Mustapha, N., Baazaoui-Zghal, H., Marie-Aude, A., and Ben-Ghzala, H.: Survey on ontology learning from web and open issues. In Third International Symposium on Innovation in Information and Communication Technology (ISI-ICT' 2009), Amman Jordan, (2009)
38. Hearst, M.A. :Automated Discovery of WordNet Relations. "Wordnet An Electronic Lexical Database". MIT Press, Cambridge, MA, 132-152, (1998)
39. Aussenac-Gilles, N., Jacques, M-P.: Designing and Evaluating Patterns for Ontology Enrichment from Texts. in proceeding of EKAW 2006, pp. 158–165, (2006)
40. Bachimont, B.: Engagement smantique et engagement ontologiqueconception et ralisation d'ontologies en Ingnierie des connaissances. Ingnierie des connaissances: volutions rcentes et nouveaux dfis, chapitre 19, (2000)
41. Faatz, A. and Steinmetz, R.: Ontology enrichment with texts from the WWW. Semantic Web Mining 2nd Workshop at ECML/PKDD-2002, Helsinki, Finland, (2002)
42. Hahn, U., Mark K.: Joint knowledge capture for grammars and ontologies. In Y. Gil, M. Musen, and J. Shavlik, editors, Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001), Victoria, British Columbia, Canada, ACM Press, pp. 68–75,(2001)
43. Hwang, CH.: Incompletely and imprecisely speaking: using dynamic ontologies for representing and retrieving information. Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99) pp. 14–20, (1999)

44. Kietz, JU., Maedche, A. and Volz, R.: A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. In: Aussenac-Gilles N, Bibow B, Szulman S (eds) EKAW'00 Workshop on Ontologies and Texts. Juan-Les-Pins, France. CEUR Workshop Proceedings,Amsterdam, The Netherlands, (2000)
45. Moldovan D. I. et Girju R. Domain-Specic Knowledge Acquisition and Classification using WordNet. In Proceeding of FLAIRS 2000 Conference, Orlando, pp. 224-228.(2000)
46. Karoui, L., Aufaure, M.-A., Bennacer, N.: Contextual Concept Discovery Algorithm, FLAIRS-20, the 20th International FLAIRS Conference, in cooperation with the American Association for Artificial Intelligence, Key West, Florida, May 7-9 2007, special track on Context in AI tools and Applications, pp. 460-465 (2007)
47. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In IJCAI, 448-453, (1995)
48. WordNet : An Electronic Lexical Database. MIT Press, (1989)
49. Sekiuchi, R., Aoki, C., Kurematsu, M., and Yamaguchi, T.: DODDLE : A domain ontology rapid development environment. Lecture Notes in Computer Science, 1531, pp. 194-204, (1998)
50. Karoui, L., Aufaure,M.-A. and Bennacer, N.: Context-based Hierarchical Clustering for the Ontology Learning, the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI-06) jointly with the 2006 IEEE/WIC/ACM International Conference on Data Mining (ICDM-06), 18-22 December 2006, Hong-Kong, IEEE Proceedings pp. 420-427, (2006)
51. Stojanovic N., Stojanovic L., Volz R.: A Reverse Engineering Approach for Migrating Data-intensive Web Sites to the Semantic Web. In: 17th World Computer Congress, Kluwer Academic Publishers, pp. 141-154, (2002)
52. Tijerino, Y. A., Embley, D. W., Lonsdale, D. W., Ding, Y., and Nagy, G. Towards ontology generation from tables. World Wide Web, Vol.8(3), 261–285, (2005)
53. Manzano-Macho, D., Gomez-Prez, A.and Borrajo, D., Unsupervised and Domain Independent Ontology Learning: Combining Heterogeneous Sources of Evidence, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), pp.28–30, (2008)
54. Allani H., Position paper: ontology construction from online ontologies, International World Wide Web Conference, pp. 491–495(2006)
55. Nakayama, K., Hara, T., and Nishio, S.: A thesaurus construction method from large scaleweb dictionaries. In AINA, IEEE Computer Society, pp. 932–939.(2007)
56. Christopher,B.: Ontology Learning from Text: Methods, Evaluation and Applications Paul Buitelaar, Philipp Cimiano, and Bernado Magnini (editors) (DFKI Saarbrcken, University of Karlsruhe, and ITC-irst), Amsterdam: IOS Press (Frontiers in artificial intelligence and applications, edited by J. Breuker et al, vol. 123, (2005)
57. Lin, D.: Automatic retrieval and clustering of similar words. Proceedings of the 17th international conference on Computational linguistics, p.768-774, August Montreal, Quebec, Canada, pp. 10–14(1998)
58. Vronis, J.: Hyperlex : lexical cartography for information retrieval, Computer Speech & Language, Vol. 18(3), pp. 223–252, (2004)

59. Navigli, R. and Velardi, P.Learning domain ontologies from document warehouses and dedicated web sites. Computational Linguistics, 30(2), 151–79, (2004)
60. Harris, Z. S. Mathematical Structures of Language. New-York, John Wiley and Sons, (1968)
61. Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science, Vol.27, 129–146, (1976)
62. Brini, A., Boughanem, M., Dubois, D.: A model for information retrieval based on possibilistic networks. In: Consens, M.P., Navarro, G. (eds.) SPIRE 2005. LNCS, vol. 3772, pp. 271282. Springer, Heidelberg (2005)
63. Velardi, P., Fabriani, P. and Missikoff, M.: Using text processing techniques to automatically enrich a domain ontology, Proceedings of the ACM Conference on Formal Ontologies and Information Systems, pp. 270–284, (2002)
64. Lebart, L., Salem, A., and Berry, L.: Exploring Textual Data. Kluwer Academic Publishers, (1998)
65. Lin, C.-Y. and Hovy, E. H. The automated acquisition of topic signatures for text summarization. In COLING, pp. 495501. Morgan Kaufmann, (2000)
66. Sanderson, M. and Croft, W.B.: Deriving concept hierarchies from text. Proceedings of the 22nd International ACM SIGIR Conference, pp. 206-213, (1999)
67. Faure D. et Nedellec, C. A corpus-based conceptual clustering method for verb frames and ontology acquisition, LREC workshop on adapting lexical and corpus resources to sublanguages and applications, Granada, Spain, (1998)
68. . Sanchez, D.: Domain ontology learning from the web. Knowledge Eng.Review, 24(4) 413, (2009)
69. Ferreira, J.: A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. World Trade, 369381, (1999)
70. Turney, P.D.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning Freiburg, Germany, pp. 491-499 (2001)
71. Downey, D., Broadhead, M., Etzioni, O.: Locating complex named entities in Web text. In Proceedings of the 20th International Joint Conference on Artificial Intelligence ,pp. 2733- 2739, (2007)
72. Lemaire, B., Denhire, G.: Effects of High-Order Co-occurrences on Word Semantic Similarities. Current Psychology Letters - Behaviour, Brain and Cognition Vol. 18(1), (2006)
73. Geleijnse, G. and Korst, J. H. M. : Automatic ontology population by googling. In Proceedings of the Seventeenth Belgium-Netherlands Conference on Artificial Intelligence, pp. 120-126, (2005)
74. Cimiano, P. Ontology learning and population from text - algorithms, evaluation and applications. Springer, (2006)
75. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A.: Web Scale Information Extraction in KnowItAll (Preliminary Results). In: Proceedings of the 13th International WWW Conference, New York, USA, pp. 100-111, (2004)
76. Kuchmann-Beauger, N. and Aufaure, M-A.: A Natural Language Interface for Data Warehouse Question Answering, 16th International Conference on Applications of Natural Language to Information Systems (NLDB 2011), June 28-30, Alicante, Spain, pp. 201-208 (2011)

77. Berners-Lee, T., Hendler, J., and Lassila, O.: The Semantic Web. Scientific American., pp. 34-43 (2001)
78. Sell, D., da Silva, D. ., Beppler, F. D., Napoli, M., Ghisi, F. B., Pacheco, R. C., and Todesco, J. L.: SBI: a semantic framework to support business intelligence. In Proceedings of the First international Workshop on ontologySupported Business intelligence (OBI). ACM,NewYork, NY,pp.111, (2008).
79. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A.: Web Scale Information Extraction in KnowItAll (Preliminary Results). In: Proceedings of the 13th International WWW Conference, New York, USA, pp. 100-111, (2004)
80. Nixon, L., Mochol, M., Jarrar, M., Dasiopoulou, S. Papastathis, V. and Kompatsiaris, Y.: Prototypical Business Use Cases. Deliverable D1.1.2 (WP1.1), The Knowledge Web Network of Excellence (NoE) IST-2004-507482, Luxemburg. January (2005)