Journal of Universal Computer Science, vol. 15, no. 4 (2009), 805-825
submitted: 30/6/08, accepted: 27/2/09, appeared: 28/2/09 © J.UCS

Fingerprinting Lexical Contexts over the Web

Vincenzo Di Lecce[*]
(Polytechnic of Bari – II Faculty of Engineering – DIASS, Taranto, Italy
v.dilecce@aeflab.net)

Marco Calabrese
(Polytechnic of Bari – II Faculty of Engineering – DIASS, Taranto, Italy
m.calabrese@aeflab.net)

Domenico Soldo
(Polytechnic of Bari – II Faculty of Engineering – DIASS, Taranto, Italy
d.soldo@aeflab.net)

Abstract: In this paper a novel technique for identifying lexical contexts in web resources is presented. The basic idea is to consider web site anchortexts as lexicalized descriptions of an individual ontology organized in the form of a graph of concept words. In the search for peculiar semantic patterns, the concept of web minutia (transposed from the forensic domain) is introduced. The proposed technique consists in searching for web minutiae in the analyzed web sites by means of a golden ontology. Web minutiae act as fingerprints for context-specific web resources; in this sense they are a powerful computational tool to identify and categorize the Web. The WordNet database has been used as golden ontology for our experiments on English web documents. WordNet allows for indexing and retrieving word senses and inter-word taxonomical relations like hyponymy and hypernymy. It has proven to be an efficient mediator between web ontologies and context-dependent taxonomies. Our experiments have been carried out on a preliminary data set of several tens of thousand links taken by web sites of thirteen UK universities. Preliminary results seem to confirm the ability of web minutiae to identify lexical contexts across the Web.

Keywords: minutia, golden ontology, Semantic Web, Web Mining, Knowledge Discovery, WordNet
Categories: L.1.4, I.2.4

1 Introduction

Since its advent, the World Wide Web (hereinafter WWW or simply the Web) has increased dramatically in size and number of interlinked resources. This trend negatively affects the precision rate of traditional search engines, which progressively lowers. The adoption of low-semantics information retrieval approaches maintains inherently the user query recall rate at low levels as anyone may experience when

[*] corresponding author

searching for a web resource specifying an ambiguous (polysemous or general purpose) query word. Consequently, search engine responses generally do not exactly match what the user actually queried for. The access to high-quality information on the Web may be thus problematic for unskilled users. The accessible part of the Web (also called *surface* Web) is then practically hidden to final user also due to the filter effect made by search engines. In addition to this, the deep (or *hidden*) Web (essentially the Web beyond dynamic HTML) is estimated to be as large as many hundred times the surface Web [Singh, 02]. It is widely accepted [Akilandeswari, 08] [Bergholz, 03] that the exploration of deep Web is partially possible only if some kind of semantic approach is used. Most popular search engines seem however to be far away from having indexed web knowledge in a semantic way. The Web itself is not structured according to semantic principles, but it resembles a heterogeneous collection of interlinked resources. Despite the multiplicity of widely recognized standards such as XML, RDF(S) [Luke, 96], DAML-OIL and OWL [Antoniou, 03], it is still too early for the Web community to adopt universal guidelines in web content publishing.

For more than a decade, research groups have been seeking for feasible ways to bridge the gap between the Web as it is and the Web as they would like it to be. In 2001, Tim Berners-Lee [Berners-Lee, 01] coined the term "Semantic Web" claiming: "*The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation*". The Semantic Web approach has the purpose of providing machines with the ability of understanding the semantic content of a website. In the W3C intentions, the Semantic Web is the place where agents operate with the tasks of: understanding the semantic content of Web pages; creating efficient routes in response to the given queries; replacing the user; creating connections among the sites and verifying the result plausibility.

Although Semantic Web still remains a chimera, countless approaches have been presented in the literature attempting to organize web knowledge from web pages. They generally rely on supervised or partially unsupervised techniques where the human factor in the knowledge engineering process is demanding. From a thorough survey on this matter in the recent literature it is worth noting that most of these approaches are taxonomy-based or ontology-based (see for example [Scime, 01] [Ganesh, 04]). However, some misconceptions often arise with the use of these terms. Nevertheless, the authors think that they represent two key concepts for clearly delimiting the span of knowledge-based techniques for the Semantic Web.

In this paper, we engage the challenge of automatically detecting lexical contexts from web resources using a golden ontology [Dellschaft, 06] as mediator between context-dependent taxonomies and ontologies represented by web pages. The basic idea is to exploit the linked structure of the Web as a basis to gain context knowledge. Web site anchortexts are here considered as lexicalized descriptions of a given context ontology organized in the form of a graph of concept words. In the search for peculiar semantic patterns, the metaphor of minutia is set forth. In the forensic domain minutiae represent the points of interest in a fingerprint that allow for disambiguating someone's identity. This is so because minutiae patterns are unique to the person. Analogously, the concept of minutia is transposed in this paper to signify special structures in the system knowledge base that help identifying context-dependent

semantic features. An introductory discussion about this topic was presented by the authors in [Di Lecce, 08a]. The proposed context-detection technique consists in searching for minutiae from the analyzed web sites. Minutiae act as fingerprints for context-specific web resources; in this sense, they are a powerful computational tool to identify and categorize the Web.

The outline of the paper is as follows: the first section is devoted to showing the advantages provided by a golden ontology. Section III presents the concept of minutia as a lexical structure to fingerprint context knowledge. Section IV deals with the implemented system used to apply the concept of minutia to the extraction of web knowledge. An extensive experiment on a group of web sites, pertaining to the same domain, is reported. Finally, conclusions and future works are drawn.

2 Representing Knowledge

In this work a graph-based model is used to represent the knowledge base (KB) provided by a lexical database (LDB) enhanced with synonymy associations. In particular, the WordNet LDB has been chosen for the analysis of English web documents. As WordNet glossary itself [Fellbaum, 98] reports, synonymy is "the semantic relation that holds between two words that can (in a given context) express the same meaning". Hence, given a word pair (x, y), an $x \rightarrow y$ assertion is interpreted as semantic relatedness between x and y in a given context. WordNet database is built around the concept of synset, that is a collection of words interchangeable within a given context. These ones are themselves related to other synsets through IS-A hierarchies [Dominguez, 06], i.e. through hyponym/hypernym relations. As Kamps [Kamps, 02] points out, WordNet implements a recursive definition of word meaning so that synonymy may be observed at different levels starting from a given concept word using IS-A hierarchies. The entire WordNet lexical taxonomy is an offspring of the super-concept "entity". Each synset accounts for the lexicalization of a concept in a given context. WordNet can be then represented in the form of a word graph where links among words play the role of synonymy relations with respect to a given context. For example, in WordNet 2.1 the query "man" produces eleven senses each of which pertains to a differently interpreted context. The assumption is made so that the chosen dictionary is strong enough to encompass all meaningful contexts, i.e. all possible senses of words are present in the dictionary itself. Although this assumption may appear crucial, it must be said that a wide effort has been already done by the WordNet community to provide a lexical reference system inspired by modern psycholinguistic theories of human lexical memory.

2.1 Taxonomy vs Ontology: Overview

The ever-growing amount of knowledge patterns scattered through the Web has been the focus of researchers' attention over the latest years. Several taxonomy and ontology-based approaches have been implemented for web knowledge categorization and retrieval.

The Oxford online dictionary[1] returns the following results for the entry word

[1] http://www.askoxford.com/dictionaries/?view=uk

"taxonomy": 1) *The branch of science concerned with classification*. 2) *A scheme of classification*. Taxonomies provide classifications among entities generally according to IS-A relations. The data model underlying taxonomy is a hierarchical structure like a tree. It seems that hierarchical relations map well to the human cognitive view of classification [Jiang, 97]. Ciaramita et al. [Ciaramita, 05] have experimentally estimated the people orientation to use specific superordinate concepts. They characterized concepts in a (lexical) taxonomy as a ranking problem and then applied ranking algorithms to evaluate the most useful superordinate concept. Obtained results seemed to testify for their assumption.

Some well-known examples of broad-coverage taxonomies can be found directly over the Web. Yahoo!, for example, is one of the first "*large taxonomy of topics: it consists of a tree of subjects, each node of which corresponds to a particular subject and is populated by relevant pages*" [Chakrabarti, 99]. More generally, it can be assumed that each web site reflects the categorization of concepts that its web designer intends to present. Such categorization is often reported in the site map as a tree-like hierarchy.

Common dictionaries generally describe the word "ontology" only through its metaphysical interpretation. The online Oxford dictionary for example reports the following definition: "*the branch of metaphysics concerned with the nature of being*". On the other hand, in computer science an ontology is defined as "*a specification of a representational vocabulary for a shared domain of discourse -- definitions of classes, relations, functions, and other objects*" [Gruber, 93]. From the definition, it follows that ontology models the world of interest, hence the *semantic layer*. Nonetheless, concepts and relations characterizing the semantic layer need to be somehow *lexicalised* in order to be expressed, then an imperative matter is on how to express specifications of concepts in a symbolic way. This is the reason why the ontological (semantic) layer requires the lexical layer. In [Su, 05] a formal specification of the two layers can be found. In other words, ontology can be considered as the formal specification for conceptualizations of a certain domain knowledge. There are some languages in the literature (OWL is one of the most popular one) that allow for describing ontologies thus making if possible for computer applications to share their KBs.

Ontologies are based on two parts, the definition of concepts and the relations among them [Gruber, 95] [Uscold, 98]. Ontology expressiveness heavily relies upon the way it is engineered. While fully automatic machine knowledge acquisition remains a chimera, some semi-automatic ontology learning approaches for the Semantic Web have been already devised by ontology engineers [Maedche, 01]. Hence, the human factor is determinant when building ontologies. An awkward aspect of this is that multiple ontologies describing the same or narrow domains may be hardly mapped each other. The same concept may be in fact lexicalised in different ways; furthermore, some relations comprised in an ontology may not be present in another one.

In a recent paper [Ning, 06] Huang Ning and Diao Shihan suggest that the structure of an ontology should satisfy the structure of its referring domain knowledge, that is the quality of the ontology strictly depends on the way its knowledge is structured. The authors regard ontology as "*an undirected graph G = <V,E>. Each concept is a vertex in this graph. If a concept has an object property whose value is an instance of another concept, an edge will be drawn between these two concepts*". Almost the

same assumptions can be found in [Chakrabarti, 99]. By the way, representing ontology knowledge in form of a graph is a widely accepted paradigm. An example in this sense is given by the RDF data model which consists in a collection of statements (each made of the triplet Subject-Verb-Object) representing a labelled directed graph [Lassila, 99]. This representation however does not include important features that a real ontology may have such as: the management of modal or fuzzy assertions, uncertainty, inconsistence and so on.

2.2 WordNet-like Golden Ontologies: Relevant Features

Both ontology-based and taxonomy-based approaches have been studied thoroughly in the literature. The ontological approaches are in general far superior than the taxonomical ones in terms of expressiveness power, but they are more difficult to implement. A system that preserves efficient categorization having also high semantic expressiveness would be the perfect compromise. In the search for a good mean between the two approaches, a midpoint seems to be represented by the new emerging golden standard-based ontologies [Zavitsanos, 08] [Ferrar, 03], which prove to be highly feasible and reliable thanks to the recent progress in developing broad-coverage dictionaries like WordNet.

Golden ontology (from now on GO for conciseness) generally cannot be in general modelled as a mere tree, because many concepts have more than one parent. Considering WordNet for example, the more appropriate data model for the hyponym/hypernym hierarchy is the directed acyclic graph (DAG) [Wagner, 04]. An indicative element that differentiates DAG from tree is that, when moving from leaves to the root in a bottom-up fashion (i.e. following hypernym patterns) there can be nodes that allow for more-than-one directions. In a lexical structure this multiplicity is due to the different senses a lexical concept may have. Contrarily to a tree, a GO in fact generally contemplates multiple contexts. Nevertheless, if more semantic relations are considered (hyponym, hypernym, meronym, holonym, antonym etc…), the DAG model may be not sufficient: in this case it is better to consider a graph model.

A GO indeed allows for browsing concepts due to its interlinked semantic relations. An immediate consequence of this is that it is possible to define semantic similarity among concepts, once defined a proper metric. GO preserves categorization but focuses on semantic capabilities. Hence, GO well interposes between the lexical layer and the semantic layer. Early considerations on these matters can be found in a work of Mejis [Meijs, 93] where the importance of semantics in machine-readable dictionaries (MRD) is clearly stated. The author stresses both the MRD navigational aspect (by proposing a "taxonomy-browser" to move within the chosen dictionary) and the MRD semantic inference aspect.

As previously reported, WordNet GO is organized around the idea of synsets, i.e. group of cognitive synonyms, each one representing a specific concept (Figure 1 reports six different senses of the concept word 'body' along with all its hyponyms). Synsets are interlinked by means of conceptual-semantic and lexical relations. For this reason, a WordNet-like database is also referred in the literature to as Lexical Knowledge Base (LKB). An LKB is a lexical resource model for indexing and retrieving word senses and inter-word relations. It is also common to find authors (for example see [Sahoo, 03]) who refer to WordNet using the locution "lexical

taxonomy". This definition is rather imprecise because it emphasizes mostly the concept categorization due to hyponym/hypernym relations.

There are several applications of WordNet spanning from word sense disambiguation [Naskar, 07] to automatic ontology extraction [An, 07]. One of the many uses of WordNet is about reckoning word similarity measures. Peng-Yuan Liu et al. [Liu, 06] present a method for assessing misaligned Chinese-English word pairs using six different pseudo-metrics. The authors use an open-source project [Newman, 03] that implements similarity computation through a Web based interface. All these metrics are based upon the concept of synset which is a group of words that share the same sense (in a given context). Distances between two synsets may be computed in several ways: the simplest one is that of considering the number of nodes along the shortest path between the synsets. This absolute value may be related to the depth of the chosen taxonomy (dictionary) in order to have a relative value (like in lch [Leacock, 98] distance measure). Wu and Palmer introduce a measure (called wup) based on the concept of Least Common Subsumer (LCS). Resnik [Resnik, 99] defines the Information Content of the LCS (see also [Kulster, 01] [Baader, 04] on this issue). Other measures have been proposed by Jiang & Conrath [Jiang, 97], Lin [Lin, 98], Banerjee and Pedersen [Banerjee, 03] etc.

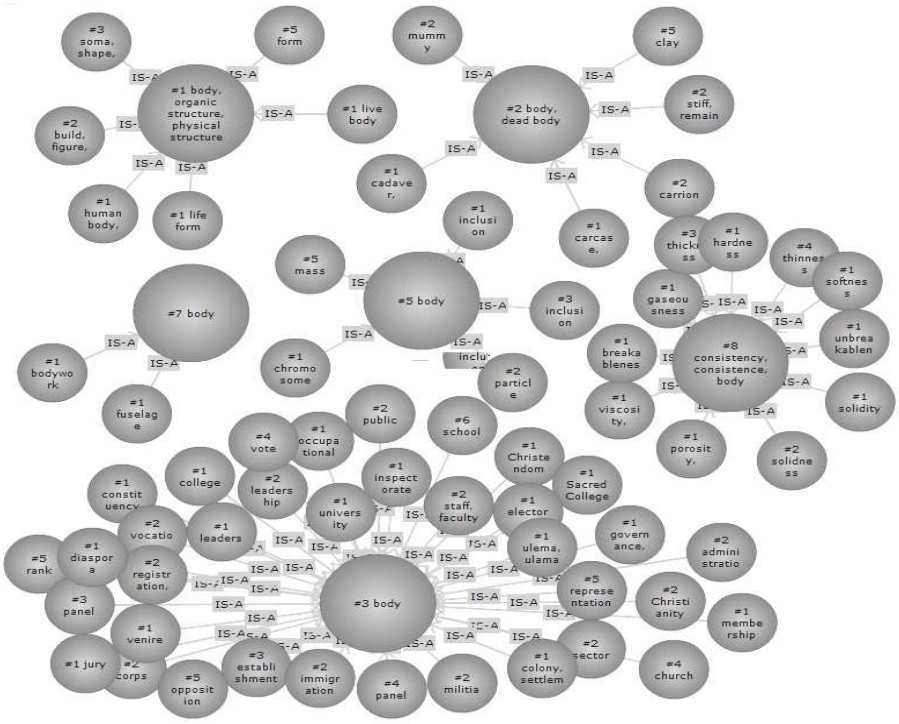

Figure 1: sample representation of the six synsets and related hyponyms containing the word concept 'body' in WordNet 2.1

Other distance measures have been proposed to assess similarity among taxonomies and ontologies (taxonomy envisages the concept classification – hierarchy – while ontology enables the formulation at complete conceptual pattern in a given website). In [Seo, 04] Seo et al. analyze the main statistical methods used to extract concept sets relevant in an ontology. The evaluation is realized by comparing the ontology obtained from each of the feature selection methods with a domain ontology manually assigned. The considered methods for the feature extraction in a dataset are: mutual information, $\chi 2$ Statistic, Markov blanket and information gain. The authors show that the mutual information and $\chi 2$ Statistic methods are better than the others; these methods are used in information retrieval for their ability to identify the words with higher semantic content for a class, respecting the ontology definition. In [Dellschaft, 06] a standard based evaluation of ontology learning by means of a lexical term analysis is proposed. The work highlights the difference between two term layers: lexical layer and hierarchical concept layer. The first layer is linked to the sets of all terms in an ontology, while the second layer depends on the semantic structure of the same ontology. The evaluation of the considered ontology (computed retrieval) is measured in comparison to a reference ontology (reference retrieval). This solution is especially suitable in broad scale evaluations and in learning methods of more ontologies.

3 "Fingerprinting" the Web

In the literature of pattern analysis ([Jain, 97] [Liang, 07]), minutiae are defined as local ridge structures (essentially endings and bifurcations) that act as local descriptors of a fingerprint. Fingerprint identification can be then reduced to the process of searching for a query set of minutiae against a given minutia database to establish the identity of an individual. Although several problems related to the matching process strictly depend on a number of variants, the idea of local descriptors characterizing an individual is attractive for web knowledge pattern discovery. Figure 2 helps visualize an example of ridge endings and bifurcation.

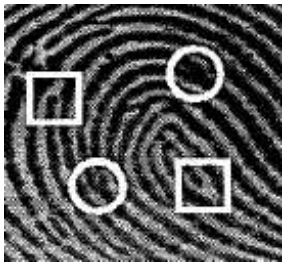

Figure 2: sample fingerprint [Prabhakar, 00]. Ridge endings are marked with white squares, while bifurcation are marked with white circles.

The attempt is made to transpose the concept of minutia to the Web for discovering relevant lexical patterns to use as *markers* of given websites. The following part will

show how to use a GO like WordNet to find out the context-dependent lexical structures that we call "web minutiae".

3.1 The Concept of Web Minutia

In this paragraph the concept of *web minutia* is introduced. The forensic concept of minutia is transposed as follows.

Let G^R and G^C be two directed node-labelled graphs:

$$G^R = \left(V^R, E^R, L_{V^R}\right) \tag{1}$$

$$G^C = \left(V^C, E^C, L_{V^C}\right) \tag{2}$$

$$L_{V^R} : V^R \rightarrow \Sigma^R,\ L_{V^C} : V^C \rightarrow \Sigma^C \tag{3}$$

with Σ^R and Σ^C representing label sets.
Assume the two graphs representing respectively the following knowledge:

- *reference* (or *golden*) ontology: plays the role of a widely accepted common ontological framework;
- *collected* ontology: plays the role of an ontology built by individuals acquired from some link-based repository like the Web.

In both graphs, nodes represent concepts while arcs account for semantic relations.
Consider the union of the two graphs $G^R \cup G^C$ and assume that vertices, having at least one label in common, represent connection points for the two graphs. In other words, the same lexical expressions define a bridge between the two ontologies. These bridge nodes can be called *homologous*. For the sake of simpleness it can be assumed that all labels on G^C nodes are also present in G^R. This limitation could be overcome by inserting any new word as representing a new synset in the chosen G^R. The following definition holds:

Def. (bifurcation and terminal nodes): $\forall e_i^R = \left(x_i^R, y_i^R\right) \in E^R$, with $x_i^R, y_i^R \in V^R$ ($x_i^C, y_i^C \in V^C$ being their homologous on G^C), $\beta^R \in V^R$ is called *bifurcation node* if $\beta^R = LCS(x_j^R, y_j^R)$; $\left(x_i^R, y_i^R\right)$ are called *terminal nodes.*

Obs. $\forall e_i^C = \left(x_i^C, y_i^C\right) \in E^R$ a bifurcation node $\beta^R \in V^R$ always exists for single-root G^R (this because LCS is a hypernym relation, hence, two nodes must have at least the root in common in a single-root taxonomy).

Def. (minutia): $\forall e_i^C = \left(x_i^C, y_i^C\right) \in E^R$ the sub-graph extracted form the two paths starting from x_i^C and y_i^C towards the bifurcation node is called minutia.

Def. (minutia order): for any given minutia, its order is defined as the number of edges encountered in the longest path from the two leaf nodes towards the bifurcation node.

An illustrative example of the previous definitions is depicted in figure 3.

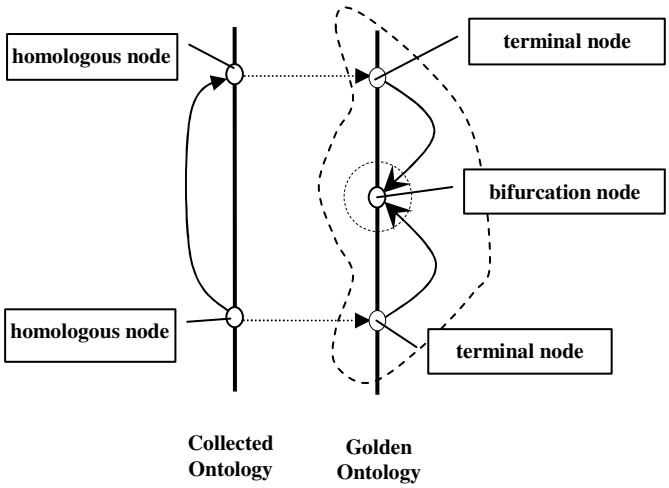

Figure 3: example of 1-order web minutia.

Minutia definition is inherently recursive: generally an *n*-order minutia can be always considered as a 1-order minutia having a 0-minutia comprising all its super-concept hierarchies as bifurcation node. 0-order minutia can be considered contemporarily as a bifurcation node and an ending node. On the other side, a minutia having leaf nodes as terminal nodes spans across the entire taxonomy, hence it has the maximal possible length for the given taxonomy. As it will be shown in the experiment paragraph, minutiae are a valid computational support to identify local structures pertaining a given context. In other words, minutiae can be used to "fingerprint" a given ontological domain. The following section presents an overview of the implemented system along with quantitative evaluations to support our assumption.

4 The Test Architecture

In order to search for web minutiae on an automated scale test architecture has been implemented using a component-based approach. A thorough system description goes beyond the scope of the work; only a brief overview is given forth.

Web minutiae extraction requires two distinct phases. Namely they are:
- (Phase I – crawling/parsing): This step consists of the user manually providing starting URIs to the system crawlers. Alternatively, the system may be also fed by URIs collected from the query result of a search engine (these URIs are generally ordered by a decreasing relevance index – e.g., the Page Rank weight calculated with the Page Rank Algorithm [Brin, 98]).
- (Phase II - semantics) The post-processing step refers to the automatic extraction of web minutiae from the document set using the knowledge provided by WordNet ([Chakrabarti, 99] i.e. our GO) and the knowledge provided by each analyzed web site (i.e. our collected ontology).

In order to carry out the two phases, the system is mainly composed of two logical components. These are:
- Knowledge Manager (KM): that handles web crawling and parsing activities as well as the analysis of the website structure.
- Web Evaluator (WE): that manages the used dictionaries, finds web minutiae and performs semantic-contextual similarity evaluation between websites.

Knowledge Manager. This component manages the operations of gaining information from web sources and storing it in a MySQL DB that can be split into three conceptually semi-independent parts:

1. Page repository: stores the analyzed web graph structure (a pointer-to-pointer table is used for this scope, that is a table that memorises the couples of adjacent links);

2. Golden ontology: the WordNet 3.0 DB is used;

3. Collected ontology: the web graph structure taken from websites is stored in accordance to graph-model described in the previous section.

The KM performs web structure mining, web content mining and their integration [Chakrabarti, 99]. Web mining activity [Kosala, 00] is generally focused on the structure and content analysis. These two issues are considered as closely related in literature and are managed by a web crawler and a parser respectively. A spider has been realised to explore the structure of a document set, whose architecture is obtained by a crawler and a page repository. KM builds the extracted ontology creating a pointer-pointer table of anchortexts taken from *adjacent* web links. We provide the following simple definition in first-order logic for adjacent web links:

Def. (adjacent Web links)
link1, link2 | ∃p, webpage(p) is_hrefof(link1,p) and has_link(p,link2))

Web Evaluator. The aim of WE is to find web minutiae and perform statistical analysis on them. First, each anchortext is tokenized and stop words are banned. Tokenization is performed according to WordNet lexical entries (words or compound words). When a word (or a compound) is found in the anchortext, it is paired with any other word (or compound) found in the adjacent link in order to have couple of words

that are semantically related trough a "href" relation. At this point, the search for web minutiae is straightforward. The component then provides for different types of statistical analysis.

4.1 Experiments and Results

Our experiments in searching for web minutiae have been carried out on a preliminary data set of tens of thousand links in the "university" domain, with particular reference to the web site of thirteen universities located in UK. Table 1 lists the targets. Our crawlers have explored 500 web pages from each web site starting from the home page according to a breadth-first crawling policy. Links outgoing the native domain have been excluded. There has been made no difference among the types of internet resources (html pages, dynamic pages, pdfs, images and so on) to crawl.

id	University name	url
1	University of Manchester	http://www.manchester.ac.uk
2	University of Birmingham	http://www.bham.ac.uk
3	University of Southampton	http://www.soton.ac.uk
4	University of Oxford	http://www.ox.ac.uk
5	Univerity of Liverpool	http://www.liv.ac.uk
6	University of Cambridge	http://www.cam.ac.uk
7	University of Warwick	http://www2.warwick.ac.uk
8	University of Brighton	http://www.brighton.ac.uk
9	University of Edinburgh	http://www.ed.ac.uk
10	University of Nottigham	http://www.nottingham.ac.uk
11	Lancaster University	http://www.lancs.ac.uk
12	Kingstone University	http://www.kingston.ac.uk
13	Loughborough University	http://www.lboro.ac.uk

Table 1: website benchmark used for web minutiae analysis

4.1.1 Method overview

The concept of web minutia does not claim to be an algorithmic solution to the general problem of knowledge extraction from the Web. Nonetheless, it represents an empirical technique that adopts a case-specific procedural approach, in order to find hidden knowledge patterns in the Web Graph (i.e. minutiae). The underpinning idea (that intercepts the trends of the recent research on the subject) is the use of a golden ontology as a "ground truth" facility in the knowledge extraction task. Words taken from parsed website anchortexts drive the search for peculiar golden ontology patterns. In this paper, the chosen golden ontology is WordNet: one of the best available semantic Lexicon of the English Language. WordNet itself is an ongoing project, since minor bugs and refinements characterize new version releases (in this work WordNet 3.0 was finally adopted).
A pseudo-code description of the implemented procedure for minutiae extraction is provided as follows:

```
% initializations
P:= list of crawled web pages
Candidates := [];      % empty list
Minutiae := [];        % empty list
c :=  -1;       % index for Candidates list
m:=  -1;       % index for Minutiae list
minutia_order := user defined;
% find candidate lexical entries from anchortexts
FOR i=0 TO length(P) - 1
   tmp1 = anchortexts of all inbound links of P[i];
   tmp2 = anchortexts of all outbound links of P[i];
   tmp_list[1] = tokenize(tmp1);
   tmp_list[2] = tokenize(tmp2);
   FOR j=0 TO length(tmp_list[1]) - 1
      FOR k=0 TO length(tmp_list[2]) - 1
         IF(exists_in_WN(tmp_list[1][j]) AND …
         exists_in_WN(tmp_list[2][k]))
            c = c + 1;
            Candidate[c][1] = tmp_list[1][j];
            Candidate[c][2] = tmp_list[2][k];
         END
      END
   END
END
% verify the presence of minutiae among candidates
FOR i=0 TO c
   % get synsets (terminal nodes) from lexical entry candidates
   tmp_synset_list[1] = get_SYNSETS(Candidate[i][1]);
   tmp_synset_list[2] = get_SYNSETS(Candidate[i][2]);
   FOR j=0 TO length(tmp_synset_list[1]) - 1
      FOR k=0 TO length(tmp_synset_list[2]) - 1
         tmp=minutia_order;
         WHILE(tmp >= 1)
            % find minutia structures starting from the candidate synsets.
            % A boolean check related to the search success is returned.
            % If check is true then returns every found minutia structure
            [check, minutia_struct_list] = find_MINUTIAE_in_WN(…
            tmp_synset_list[1][j],tmp_synset_list[2][k], tmp);
            IF(check)
               FOR t=0 TO length(minutia_struct_list) - 1
                  m = m + 1;
                  % retrieve and store all synsets encountered in paths
                  % from the leaf nodes towards the bifurcation node
                  Minutiae[m].synset=[get_SYNSETS_from_STRUCT(…
                        minutia_struct_list[t].path1)…
                        get_SYNSETS_from_STRUCT(…
                        minutia_struct_list[t].path2)…
                        get_SYNSETS_from_STRUCT(…
```

```
49.                                    minutia_struct_list[t].bifurcation_node)];
50.                 END
51.             END
52.             tmp = tmp - 1;
53.         END
54.     END
55. END
56. END
```

4.1.2 Used metrics

Two well-known metrics have been adapted for quantitative purposes: Lexical Recall (LR) and Lexical Precision (LP). Although there may be found many other ontology learning methods using a GO in the very recent literature [Zavitsanos, 08], LP and LR may be considered the basic measures for their simpleness and efficacy. Using the formalism in [Cimiano, 05], LR and LP have been renamed as Synset Lexical Recall (SLR) and Synset Lexical Precision (SLP) this way:

$$SLP(O_1, O_2) = \frac{|C_1 \cap C_2|}{|C_2|} \tag{4}$$

$$SLR(O_1, O_2) = SLP(O_2, O_1) \tag{5}$$

Similarly, the harmonic mean is computed as follows:

$$F(O_1, O_2) = \frac{2 \times SLR(O_1, O_2) \times SLP(O_1, O_2)}{SLR(O_1, O_2) + SLP(O_1, O_2)} \tag{6}$$

All these measures range from 0 to 1. C_1 and C_2 ($C_1 \equiv V^{O_1}, C_2 \equiv V^{O_2}$) represent the set of concepts characterizing the two ontologies O_1 and O_2. O_1 and O_2 are the ontologies extracted from each web site (collected ontologies) according to web minutia definition. In our case C_1 and C_2 correspond to the set of synsets that belongs to found web minutiae. It is noteworthy that if WordNet is considered as the golden ontology C_{golden} will result:

$$C_i \subseteq C_{golden} \forall i \mid O_i \text{ is an extracted ontology}$$

4.1.3 Data analysis

For the sake of clarity the analysis of the obtained data has been split into three sequential steps which are described as follows.

Step 1 – Web minutiae extraction. According to the given definition for web minutia, the set of web minutiae of each web site has been extracted. 0 to 2-order web minutiae have been reckoned. Higher depths require more computational time since the implemented graph-search technique is based on a Dijkstra algorithm requiring

$O(n^2)$ time. The enhancement of the search technique is currently under way. From the whole amount of web minutiae pertaining to each web site, involved synsets have been considered to assess $|C_i|$ values. The number of minutiae found for each web site is shown in Table 2.

	#minutiae			
id	**0-order**	**1-order**	**2-order**	**TOT**
1	10	16	67	**93**
2	6	22	127	**155**
3	10	32	327	**369**
4	5	16	79	**100**
5	7	34	328	**369**
6	9	32	137	**178**
7	5	18	94	**117**
8	10	10	366	**386**
9	6	40	369	**415**
10	9	73	386	**468**
11	7	28	273	**308**
12	5	72	165	**242**
13	3	25	534	**562**

Table 2: minutiae retrieved through web minutiae extraction from the benchmark

Each minutia accounts for one or many synsets depending on its order and on the number of senses one minutia node may have. It may also happen that two or more minutiae share part of their synsets. Experimentally it was found out that generally there is a small set of synsets with high occurrence while the others degrade to low values. Figure 4 illustrates the number of benchmark synsets plotted at decreasing occurrence values.

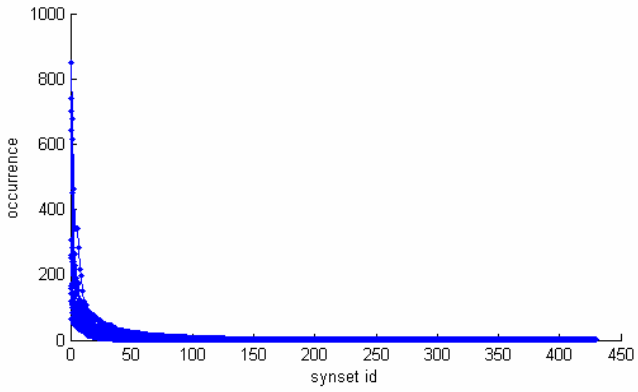

Figure 4: benchmark synset occurrences in decreasing order

Step 2 – Websites in comparison. After $|C_i|$ computation, SLP, SLR and F are straightforward to obtain. However, as pointed out before, low-occurrence synsets represent outliers with respect to the core semantics of the analyzed web sites. Hence a threshold on synset occurrence has been implemented to reduce the number of considered synsets to the most representative ones. Figure 5 illustrates maximum SLP, SLR and F-measure values obtained at increasing threshold values. It is noteworthy that SLP (SLR) peaks unity at about 60 threshold value. To have an immediate look of the total benchmark, a grayed image has been taken from the following matrices:

$$M_{SLP} = \left(SLP(O_i, O_j)\right), \; M_{SLR} = \left(SLR(O_i, O_j)\right), \; M_F = \left(SLR(O_i, O_j)\right)$$

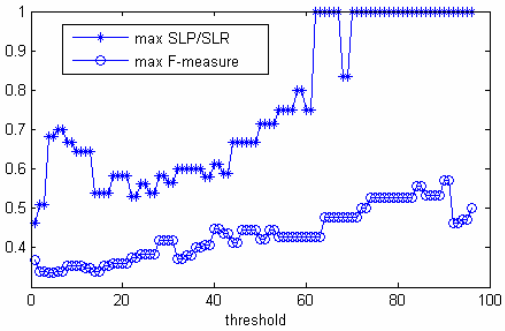

Figure 5: max values of SLP/SLR and F-measure at different threshold values

Figure 6 depicts these matrices taken at 100 threshold value. Results seem to confirm that there are implicit common lexical structures among different web sites. The authors think that this is of great interest for web site categorization and in general for the Semantic Web.

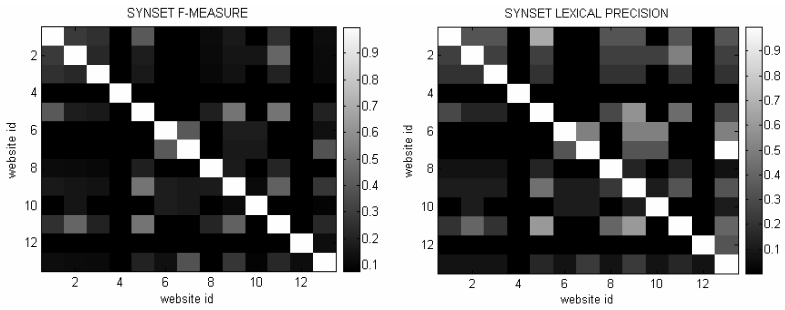

Figure 6: grayed image representing synset F-measure and SLP/SLR over the entire set of analyzed web sites

Step 3 – From individual (subjective) to common knowledge organization. By construction, web minutiae help finding a common lexical structure (WordNet) among ontologies provided by individuals (websites). Although finding the more appropriate metric for evaluation purpose may be a great deal, it can be useful to look at some practical example for having a clear mind of how web minutiae actually work. In particular one of the many hundreds of web minutiae found are reported: 1-order web minutia having the concept lexicalized with the term 'body' as bifurcation node. Table 3 lists all circumstances under which this web minutia was found. A pictorial representation of table data is depicted in Figure 7.

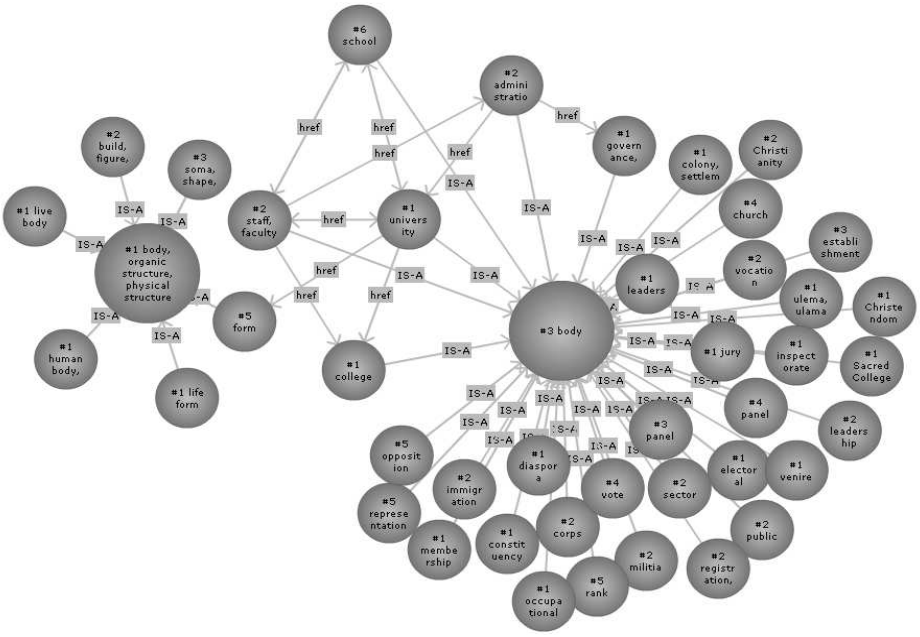

Figure 7: Pictorial representation of data in Table 3. The word concept 'body' in WordNet 2.1 is interleaved with knowledge extracted from web sites, according to web minutia definition.

It is noteworthy that different web sites (i.e. different individual ontologies) use different concept words (through the anchortexts of their web pages) to mean the same super-concept. A pure lexical system like a text-based search engine would not retrieve any relation of this kind among different websites. Furthermore, it is surprising that all but one web minutiae have the same sense (out of six) for 'body'. This means that the editors of that websites were sharing the same linguistic concept when wrote down those pages.

Bifurcation node	Terminal node (branch 1)	Terminal node (branch 2)	Website id
'body'#3	'university'	'staff'	2
'body'#3	'staff'	'college'	4
'body'#3	'university'	'college'	4
'body'#3	'college'	'university'	6
'body'#3	'administration'	'governance'	9
'body'#3	'administration'	'university'	9
'body'#3	'school'	'faculty'	10
'body'#3	'school'	'staff'	10
'body'#3	'school'	'university'	10
'body',#3	'university'	'school'	10
'body',#3	'staff'	'school'	11
'body'#1,#3	'university'	'form'	12
'body',#3	'staff'	'faculty'	13
'body',#3	'staff'	'organisation'	13

Table 3: distribution of 1-order web minutia 'body' across web sites

4.1.4 Overall considerations

The underpinning idea in the hereby proposed concept of minutia is the attempt at finding markers that characterize the system KB locally. In our case, the system knowledge base is the union of a golden lexical-semantic ontology like WordNet with the individual ontology extracted from webpages. At some extent, this approach is quite close to other recent ones like YAGO ontology [Suchanek, 07] that uses a combination of rule-based and heuristic methods to synthesize knowledge extracted from Wikipedia with WordNet database. Our paper however focuses more on the developed information extraction technique. At the same time, the idea of minutia follows the footprints of latest research upon lexical chains (see for example [Doran, 04]) which seems to be a promising one in the literature.

The here presented minutiae-based technique has several aspects that still need to be dealt with in our prospective research. First of all, a deeper understanding of the relations (if any) between lexical context and semantic domains should be investigated. Late works on the subject [Bentivogli, 04] attempt to label each synset in WordNet with domain descriptors. This assumption could be tested using the proposed technique. It is worth stressing that our method strongly depends on the chosen golden ontology, this means that WordNet, as it is, may be not a sufficient vocabulary for narrow domains. Furthermore, other interesting features of WordNet like the rank assigned to senses (depending on their occurrence in corpora) may be used for weighing synsets retrieved from the minutiae extraction task. Another important issue is that anchortexts could be enriched with other lexical descriptions extracted from the webpage. Actually, our simplifications that reduce the Web Graph to a Word Graph with texts only taken from links was a work hypothesis to simplify WWW scanning and to reduce memory storage.

In our future work the plan is to extend minutiae finding to other-than IS-A relations. It would be also interesting to assess the statistical relation between webpages producing minutiae and their minutia order. The application of minutiae for semantic tagging and agent-based systems is currently under way [Di Lecce, 08b].

5 Conclusions

In this paper a novel technique for automatically distinguishing web resources by means of their lexical contexts has been presented. The idea was to consider the lexical structures derived from the Web-graph as a fingerprint. As it happens for the forensic domain, a fingerprint always has specific structures called minutiae that help understand the identity of the person one wants to discover. In the case of web resources minutiae are lexical structures that can be extracted from individual ontologies (like web sites) by means of a golden ontology like WordNet. A rich set of experiments has been carried out in order to test the efficacy of the proposed method. Lexical precision and recall index have been computed for this scope. First results seem promising. Web minutia can be a powerful instrument to capture the lexical structures that characterize individual ontologies. In the next future the goal will be to enhance our formalism and extend these evaluations to other well-known measures. Other future efforts will be devoted to the evolution of golden knowledge construction by means of artificial intelligence systems, the creation of post-processing refinement morphological methods, the realization of a finished version of a semantic search engine based on the proposed techniques, the exploitation of partially unsupervised intelligent crawling and parsing techniques. Finally, the generality of the approach make it suitable for other application domains where knowledge can be represented in form of graph of word concepts.

References

[Akilandeswari, 08] Akilandeswari, J., Gopalan, N. P.: “An Architectural Framework of a Crawler for Locating Deep Web Repositories using Learning Multi-agent Systems”; Proc. of the 3rd International Conference on Internet and Web Applications and Services (ICIW 08), pp. 558 – 562, June 8-13, 2008, Athens, Greece.

[An, 07] An, Y. J., Geller, J., Wu, Y. T., Chun, S. A.: “Automatic Generation of Ontology from the Deep Web”; 18th International Workshop on Database and Expert Systems Applications (DEXA 07), pp. 470 – 474, September 3-7, 2007, Regensburg, Germany

[Antoniou, 03] Antoniou, G., van Harmelen, F.: “Web Ontology Language: OWL”; in Staab S. and Studer R., Handbook on Ontologies in Information Systems, Springer-Verlag, 2003, pp. 76 – 92.

[Baader, 04] Baader, F., Sertkaya, B., Turhan, A.Y.: “Computing the Least Common Subsumer w.r.t. a Background Terminology”; Proc. of the 9th European Conference on Logics in Artificial Intelligence (JELIA 04), September 27-30, 2004, Lisbon, Portugal. Published in Lecture Notes in Computer Science, Vol. 3229/2004, pp. 400 – 412, Springer Berlin/Heidelberg.

[Banerjee, 03] Banerjee, S., Pedersen, T.: "Extended Gloss Overlaps as a Measure of Semantic Relatedness"; Proc. of the 18th International Joint Conference on Artificial Intelligence, (IJCAI, 03), pp 805 – 810, August 9-15, 2003, Acapulco, Mexico.

[Bentivogli, 04] Bentivogli, L., Forner, P., Magnini, B., Pianta, E.: "Revising The Wordnet Domains Hierarchy: Semantics Coverage And Balancing"; MLR2004: PostCOLING Workshop on Multilingual Linguistic Resources, pp. 94 – 101, August 28, 2004, Geneva, Switzerland.

[Bergholz, 03] Bergholz, A., Childlovskii, B.: "Crawling for domain-specific hidden Web resources"; Proc. of the 4th International Conference on Web Information Systems Engineering, (WISE, 03), pp. 125 – 133, December 10-12, 2003, Rome, Italy.

[Berners-Lee, 01] Berners-Lee, T., Hendler, J., Lassila, O.: "The Semantic Web"; Scientific American, May, 2001.

[Brin, 98] Brin, S., Page, L.: "The Anatomy of a Large-Scale Hypertextual Web Search Engine"; Computer Networks and ISDN Systems, Vol. 30, No. 1-7, pp. 107 – 117, 1998.

[Chakrabarti, 99] Chakrabarti, S., Dom, B. E., Gibson, D., Kleinberg, J. M., Kumar, R. S., Raghavan, P., Rajagopalan, S., Tomkins, A.: "Mining the Link Structure of the World Wide Web"; IEEE Computer, Vol. 32, No. 8, pp. 60 – 67, 1999.

[Ciaramita, 05] Ciaramita, M., Sloman, S., Johnson, M., Upfal, E.: "Hierarchical Preferences in a Broad-Coverage Lexical Taxonomy"; Proc. of the 27th Annual Conference of the Cognitive Science Society, (CogSci, 05), pp. 459 – 464, July 21-23, 2005, Stresa, Italy.

[Cimiano, 05] Cimiano, P., Hotho, A., Staab, S.: "Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis"; Journal of Artificial Intelligence research, Vol 24, pp. 305 – 339, 2005.

[Dellschaft, 06] Dellschaft, K., Staab, S.: "On How to Perform a Gold Standard Based Evaluation of Ontology Learning"; Proc. of the 5th International Semantic Web Conference, (ISWC, 06), pp. 173 – 190, Athens, GA, USA, November 5-6, 2006.

[Di Lecce, 08a] Di Lecce V., Calabrese M., Soldo D.: "Mining Context-Specific Web Knowledge: an Experimental Dictionary-based Approach"; Proc. of the International Conference on Intelligent Computing, (ICIC, 08), September 15-18, 2008, Shanghai, China. Published in Lecture Notes in Computer Science, Vol. 5227/2008, pp. 896 – 905, Springer Berlin/Heidelberg.

[Di Lecce, 08b] Di Lecce, V., Calabrese, M.: "Taxonomies and Ontologies in Web Semantic Applications: the New Emerging Semantic Lexicon-Based Model"; IEEE International Conference on Intelligent Agents, Web Technologies and Internet Commerce, (IAWTIC, 08), December 10-12, 2008, Vienna, Austria.

[Dominguez, 06] Domınguez, E., Lloret, J., Rubio, A. L., Zapata, M. A.: "Evolving the Implementation of ISA Relationships in EER Schemas"; Proc. of the Workshop on Evolution and Change in Data Management, (ECDM, 06), November 6-9, 2006, Tucson, AZ, USA. Published in Lecture Notes in Computer Science, Vol. 4231/2006, pp. 237 – 246, Springer Berlin/Heidelberg.

[Doran, 04] Doran, W., Stokes, N., Carthy, J., Dunnion, J.: "Comparing lexical chain-based summarisation approaches using an extrinsic evaluation"; Proc. of the 5th International Conference on Intelligent Text Processing and Computational Linguistics, (CICLing, 04), pp. 112 – 117, February 15-21, 2004, Seoul, Korea.

[Fellbaum, 98] Fellbaum, C.: "WordNet: An electronic lexical database"; MIT Press, Cambridge, May, 1998.

[Ferrar, 03] Farrar, S., Langendoen, D. T.: "A linguistic ontology for the Semantic Web"; GLOT International Vol. 7, No. 3, pp. 97 – 100, March, 2003.

[Ganesh, 04] Ganesh, S., Jayaraj, M., Kalyan, V., Murthy, S., Aghila, G.: "Ontology-based Web Crawler"; Proc. of the International Conference on Information Technology: Coding and Computing, (ITCC, 04), pp. 337 – 341, April 5-7, 2004, The Orleans, Las Vegas, Nevada, USA.

[Gruber, 93] Gruber, T. R.: "A Translation Approach to Portable Ontology Specifications"; Knowledge Acquisition, Vol. 5, No. 2, pp. 199 – 220, June, 1993.

[Gruber, 95] Gruber, T. R.: "Toward principles for the design of ontologies used for knowledge sharinge"; International Journal of Human and Computer Studies, Vol. 43, No. 5-6, pp. 907–928, November/December, 1995.

[Jain, 97] Jain, A. K., Hong, L., Bolle, R.: "On-Line Fingerprint Verification"; IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 4, pp.302 – 314, April, 1997.

[Jiang, 97] Jiang, J. J., Conrath, D. W.: "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy"; Proc. of International Conference on Research on Computational Linguistics (ROCLING X), pp. 19 – 33, August 22-24, 1997, Taipei, Taiwan.

[Kamps, 02] Kamps, J.: "Visualizing WordNet Structure"; 1st International Wordnet Conference (GWC, 02), pp. 182 – 186, January 21-25, 2002, Mysore, India.

[Kosala, 00] Kosala, R., Blockeel, H.: "Web Mining Research: A Survey"; ACM SIGKDD Explorations Newsletter, Vol. 2, No. 1, pp. 1 – 15, June, 2000.

[Kulster, 01] Kulster, R., Borgida, A.: "What is an Attribute? Consequences for the Least Common Subsumer"; Journal of Artificial Intelligence Research, Vol. 14, pp. 167 – 203, 2001.

[Lassila, 99] Lassila, O., Swick. R.: "Resource Description Framework (RDF) Model and Syntax Specification"; The World Wide Web Consortium (W3C), W3C Recommendation, February 22, 1999.

[Leacock, 98] Leacock, C., Chodorow, M.: "Combining local context and WordNet similarity for word sense identification"; An Electronic Lexical Database, 1998, pp. 265 – 283.

[Liang, 07] Liang, X., Bishnu, A., Asano, T.: "A Robust Fingerprint Indexing Scheme Using Minutia Neighborhood Structure and Low-Order Delaunay Triangles"; IEEE Transactions on Information Forensics and Security, Vol. 2, No. 4, pp. 721 – 733, December, 2007.

[Lin, 98] Lin, D.: "An information-theoretic definition of similarity"; Proc. of the 15th International Conference on Machine Learning, (ICML, 98), pp. 296 – 304, July 24-27, 1998, Madison, WI, USA.

[Liu, 06] Liu, P.Y., Zhao, T.J., Yu, X. F.: "Application-Oriented Comparison and Evaluation of Six Semantic Similarity Measures Based on Wordnet"; Proc. of the 5th International Conference on Machine Learning and Cybernetics, (ICMLC, 06), pp. 2605 – 2610, August 13-16, 2006, Dalian, China.

[Luke, 96] Luke, S., Specter, L., Rager, D.: "Ontology-based knowledge discovery on the World Wide Web"; Working Notes of the Workshop on Internet-Based Information Systems at the 13th National Conference on Artificial Intelligence (AAAI, 96), pp. 96 – 102, August 4-6, 1996, Portalnd, Oregon, USA.

[Maedche, 01] Maedche, A., Staab, S.: "Ontology Learning for the Semantic Web"; IEEE Intelligent Systems, Volume 16, Issue 2, pp. 72 – 79, March 2001.

[Meijs, 93] Meijs, W.: "Inferring grammar from lexis: machine-readable dictionaries as sources of wholesale syntactic and semantic information"; IEEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, pp. P3/1 – P3/5, April, 1993.

[Naskar, 07] Naskar, S. K., Bandyopadhyay, S.: "Word Sense Disambiguation Using Extended WordNet"; Proc. of the 17th International Conference on Computing: Theory and Applications, (ICCTA, 07), pp. 446 – 450, March 2-7, 2007, Kolkata, India.

[Newman, 03] Newman, M. E. J.: "The Structure and Function of Complex Networks"; SIAM Review, Vol. 45, No. 2, pp. 167 – 256, 2003.

[Ning, 06] Ning, H., Shihan, D.: "Structure-Based Ontology Evaluation"; Proc. Of the International Conference on e-Business Engineering, (ICEBE, 06), pp. 132 – 137, October 24-26, 2006, Shanghai, China.

[Prabhakar, 00] Prabhakar, S., Jain, A. K., Wang, J., Pankanti, S., Bolle, R.: "Minutia Verification and Classification for Fingerprint Matching Proceedings"; Proc. of the 15th International Conference on Pattern Recognition, (ICPR, 00), pp. 25 – 29, September 3-8, 2000, Barcelona, Spain.

[Resnik, 99] Resnik, P.: "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Applications to Problems of Ambiguity in Natural Language"; Journal of Artificial Intelligence Research, Vol. 11, pp. 95 – 130, 1999.

[Sahoo, 03] Sahoo, K., Vidyasagar, V. E.: "Kannada WordNet - A Lexical Database"; Proc. of the Conference on Convergent Technologies for Asia-Pacific Region (TENCON, 03), pp. 1352 – 1356, October 15-17, 2003, Bangalore, India.

[Scime, 01] Scime, A., Kerschberg, L.: "WebSifter: An Ontological Web-Mining Agent for E-business"; Proc. of the 9th Working Conference on Database Semantics: Semantic Issues in E-Commerce,(IFIP TC2/WG2.6), pp. 187 – 201, April 25-28, 2001, Hong Kong, China.

[Seo, 04] Seo, Y. W., Ankolekar, A., Sycara, K.: "Feature Selection for Extracting Semantically Rich Words"; Technical report CMU-RI-TR-04-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, March, 2004.

[Singh, 02] Singh, M. P.: "Deep Web Structure"; IEEE Internet Computing, Vol. 6, No. 5, pp. 4 – 5, September, 2002.

[Suchanek, 07] Suchanek, F. M., Kasneci, G., Weikum, G.: "Yago: a core of semantic knowledge"; Proc. of the 16th International Conference on World Wide Web (WWW, 07), pp. 697 – 706, May 8-12, 2007, Banff, Alberta, Canada.

[Su, 05] Su, C., Gao, Y., Yang, J., Luo, B.: "An Efficient Adaptive Focused Crawler Based on Ontology Learning"; Proc. of the 5th International Conference on Hybrid Intelligent Systems (HIS, 05), pp. 73 – 78, November 6-9, 2005, Rio de Janeiro, Brazil.

[Uscold, 98] Uschold, M., King, M., Moralee, S., Zorgios, Y.: "The enterprise ontology"; The Knowledge Engineering Review, Vol. 13, pp. 31 – 89, 1998.

[Wagner, 04] Wagner, A.: "Estimating Frequency Counts of Concepts in Multiple-Inheritance Hierarchies"; LDV Forum, Vol. 19, pp. 81 – 91, 2004.

[Zavitsanos, 08] Zavitsanos, E., Paliouras, G., Vouros, G. A.: " A Distributional Approach to Evaluating Ontology Learning Methods Using a Gold Standard"; Proc. of 3rd Workshop on Ontology Learning and Population (OLP3) at ECAI 2008, July 21-22, 2008, Patras, Greece.