On Designing Semantic Lexicon-Based Architectures for Web Information Retrieval

Vincenzo Di Lecce, Marco Calabrese
DIASS
Politecnico di Bari – II Faculty of Engineering
Taranto, Italy
e-mail: {v.dilecce, m.calabrese}@aeflab.net

Domenico Soldo
myHermes S.r.l.
Taranto - Italy
e-mail: domenico.soldo@myhermessrl.com

***Abstract*—In this work, a novel framework for designing Web Information Retrieval systems with particular reference to semantic search engines is presented. The key idea is to add the semantic dimension to the classical Term-Document Matrix thus having a three-dimensional dataset. This enhancement allows for defining a lexico-semantic user interface where the query process is performed at the conceptual level thanks to the use of a Semantic Lexicon. WordNet Semantic Lexicon is used here as golden ontology for handling polysemy and synonymy, hence it is useful for disambiguating user queries at the semantic level. A layered multi-agent system is employed for supporting the design process. Particular emphasis is given to formal system knowledge representation, the interface layer managing user-system interaction and the markup layer performing the semantic tagging process.**

Keywords-component; information retrieval; semantic lexicon; WordNet; MAS; semantic query

I. 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 enforces search engine developers to adopt Web document indexing techniques, which exchange scalability for fair precision/recall performances. Surveying the literature of the latest years it is easy to notice the growing consensus about the need for involving semantics in retrieval systems. Approaches that employ low-level features as indexing parameters are prone in fact to a number of pitfalls, like the inherent ambiguity of polysemous query words. At the present time, commercial search engines provide relevant responses if the user is good enough when submitting the *right* query. Therefore, the access to high-quality information on the Web may be still problematic for unskilled users.

Traditional search engines are conceptually based on a term-document look-up table (also known as Term-Document Matrix or TDM for short). Lexical terms conveyed by the user query play the role of entries; documents populate a (ranked) list of weblinks that match user query terms according to a given metric. The user is required to discern among the given options and choose the one that is supposed to be closest to his/her intentions.

A more sophisticated type of Web information retrieval systems is represented by meta-search engines, which relay user query to several search engines, collect their responses and finally propose them to the user according to certain criteria. Meta-search engines, however, have still to deal with the problem of mixing information coming from different sources, which is an awkward task to accomplish, unless some semantic approach is pursued.

Semantic search engines should attempt to understand the user query at the ontology level. They should also offer a pictorial representation of the retrieved dataset, letting the user have the impression to move within a semantic search space. In order to be really effective, they require a strong theoretical knowledge model, which has to be sufficiently robust to allow for indexing heterogeneous data scattered across the Web.

In this work, a novel framework for designing textual information retrieval systems with particular reference to semantic search engines is presented. A semantic dimension is added to the classical term-document matrix thus having a three-dimensional dataset. In this view, the user is forced to adopt a new semantic query paradigm, which is closer to human understanding than to traditional keyword-based techniques. The query process is performed at the conceptual level thanks to the use of a Semantic Lexicon considered as a golden ontology useful for the sense disambiguation task. A layered Multi-Agent System (MAS) is employed for supporting the whole design process.

This article is an extension of a previous work presented in the ICIW 2009 Conference [1] specifically focused on semantic tagging of Web resources using MAS architecture. In the present paper, the critical point of including semantics in text retrieval systems is handled under a more general and complete perspective that involves Web ontology modeling. The final aim is to bridge the gap between traditional search engines based on term-document indexing and emerging semantic requirements by means of a suitable model, which embeds terms, documents and semantics into a single knowledge representation.

The outline of the paper is as follows: Section II reports related work in Information Retrieval with particular reference to search engines and semantic tagging aspects; Section III describes WordNet architecture [2] and its usefulness for the scope of this work; Section IV proposes the new three-dimensional information retrieval framework; Section V presents the used multi-agent system architecture; Section VI comments the carried out experiments and prototypal implementations; conclusions are sketched in Section VII.

II. Related Work

Information Retrieval (IR) is finding material (usually documents) of unstructured nature (usually text) satisfying an information request from within large collections (usually stored on computers) [3]. Automated IR systems are conceptually related to object and query. In the context of IR systems, an object is an entity, which keeps or stores information in a database, i.e. in a structured repository. User queries are then matched to objects stored in the database. A document is, therefore, an opportune collection of data objects.

Often the documents themselves are not kept or stored directly in the IR system, instead they are represented in the system by document surrogates automatically generated by the same IR system by means of a document analysis. Nowadays there are two approaches to document analysis: statistical and semantic.

The statistical approach was initially proposed by Lhun. In 1958 he wrote: "*It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements*" [4]. It is interesting to note that this approach is still used in many modern IR systems.

On the other hand, a Semantic Information Retrieval system exploits the notion of *semantic similarity* (based on lexical and semantic relations) between concepts to determine the relevancy of a certain document. One way of incorporating semantic knowledge into a representation is mapping document terms to ontology-based concepts. In [5], for example, a formal ontology-based model for representing Web resources is presented. Starting from semantic Web standards as well as established ontologies the authors reformulate the IR task into a data retrieval task assuming that more expressive resources and query models allow for a precise match between content and information needs. In this work instead, the term-concept mapping is provided by a golden ontology expressed in the form of a Semantic Lexicon like WordNet. The usefulness of this choice will be explained throughout the text further on.

A. Traditional IR techniques

The most widespread and popular applications of IR are Web search engines. They are designed to answer to a human query with an HTML page containing a ranked list of links to Web sites or documents. Every traditional Web search engine represents each retrieved webpage in its own search space by using a set of sentences that are considered as relevant to the user query. The relevancy of the retrieved documents is essentially dependent upon the chosen metric and the ranking strategy. As far as now, the most common document retrieval approach is searching for word-to-word correspondences (after stemming and stop-word procedures) between the set of query keywords and the set of document terms. Although the query search may be restricted by using Boolean and/or operators (thus providing a more selective filtering on the search space), the quality of document retrieval is significantly affected by the ranking strategy. A simple comparison among the principal Web search engines shows in fact how different the retrieved document could be, even in response to the same user query word.

The well-known Page Rank Algorithm [6] has been one of the keys to success for the Google Web search engine. It represents undoubtedly one of the most single important contributions to the field of IR in the latest years. The Page Rank Algorithm employs a fast convergent and effective random-walk model for ranking graph nodes like hyperlinked Web resources [7]. It is based on the bright assumptions that weblinks may be interpreted as "votes" given from the source page to the destination page. The vote expressed by a link is in fact weighted by the "reference" (Page Rank value) of the pages from where the links come, in accordance with the formula provided by the authors:

$$PR(A) = (1-d) + d * \left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right) \quad (1)$$

where *PR(X)* function gives the Page Rank value of page *X*, *A* is the webpage pointed by T_1, T_2,... T_n webpages, $C(T_i)$ is the number of links outgoing from page T_i and finally *d* is a properly set constant value. The previous formula is recursive. By highly ranking the most referenced pages, Page Rank represents a good prior filter to the enormous heterogeneous search space. In addition to this, the simple graphic view provided by Google home page can be easily understood by a great variety of users. In many real cases, Google apparent precision, however, can be partially ascribed to the poor syntax underpinning the user query and to the self-influence it has had on users in the way they formulate the query. Everyone can experience how much the retrieval performances decrease with more complex human-like queries.

The adoption of more sophisticated retrieval functions can help reduce the misbalance now pending on the ranking algorithms.

B. Semantic IR techniques

To overcome the limits of the traditional approaches, new semantics based techniques are being investigated in the latest years, although there is no ground-breaking technology at the moment that can be considered sufficiently mature to compete with traditional IR systems on a large scale. In the preface to the proceedings of a late international workshop on semantic search held in 2008 [8] it is explicitly stated: *"...the representation of user queries and resource content in existing search appliances is still almost exclusively achieved by simple syntax-based descriptions of the resource content and the information need such as in the predominant keyword-centric paradigm."*. A recent study [9] shows that retrieval performances are still low for both keyword-based search engines and the semantic search engines.

Provided this, one of the most relevant semantic techniques which has had a number of useful applications in various fields spanning from information discovery to document classification is Latent Semantic Indexing (LSI). LSI implements a strictly mathematical approach based on applying Single Value Decomposition (SVD) to the TDM. SVD decomposes the TDM into the product of three matrices:

$$TDM = T_0 * S_0 * D_0' \qquad (2)$$

T_0 and D_0 are the matrices of left and right singular vectors and S_0 is the diagonal one with its elements representing singular values in decreasing order. D_0' is the matrix transpose of D_0. Taking only the largest values offers a good approximation of the original TDM, thus reducing the whole search space to a relatively smaller "concept space" called LSI [10]. From this point of view, LSI is more powerful than a traditional document search algorithm: it overcomes the limits of Boolean query allowing for clustering documents semantically. LSI represents a right compromise between simplicity and good retrieval performance (measured as recall/precision values). This makes it a powerful, generic technique able to index any cohesive document collection in any language. It can be used in conjunction with a regular keyword search, or in place of one, with good results. Unfortunately, LSI suffers from scalability problems since large document sets require heavy computing on massive matrices. Furthermore, although it has been shown that LSI is able to handle correctly data structured into taxonomic hierarchies [11], it is not suited to make these taxonomies explicit in the search results. In other words, LSI is a good tool for finding semantic similarities, but it clusters output data in a *flat* (nonhierarchical) manner.

To deal with both semantics and lexical issues, a more comprehensive approach than TDM-based techniques is needed. This work grounds on the idea that, for an efficient information retrieval, lexical forms must be endowed with semantic tags in order to disambiguate their meaning. To carry out the disambiguation task, WordNet Semantic Lexicon is used.

C. Semantic Tagging

Semantic tagging (or markup) is conceived to define metadata for describing a given resource. A tag can be interpreted as a placeholder that helps user (human or computer) understands the context in which to interpret the tagged resource. An HTML tag, for example, lets browser interpret how to render a webpage; an XML tag instead allows for defining an entity name in a syntactically structured way. However, despite the initial enthusiasm around this new (meta)language [12], XML alone proved to be insufficient for most ontology-driven applications. XML in fact supplies a well-defined syntax (which is a desirable for data integration) but lacks in providing semantics. For example, XML does not resolve the lexical ambiguity that may arise when two applications share data having the same tag names, unless a Document Type Definition (DTD) or a XML Schema Definition (XSD) file is attached.

The authors are confident that any kind of semantic application cannot exist without prior defining the knowledge representation model, which is suitable and sufficient to express the given problem ontology. In the semantic engineering process, the ontology that conceptualizes (ideally in the best way) the common body knowledge is generally called *golden* or *gold standard* ontology. Its counterpart is the *individual* ontology, which strongly depends on the person who actually performs the ontology engineering process.

Generally, Web ontology modeling requires an engineering effort that can be yielded only by experts with the aid of auxiliary ontology editing tools [13][14]. In the last decade much attention has been devoted to designing layered XML-based languages such as RDF(S) [15], DAML-OIL and OWL [16], all based on formal semantics. The final attempt was to find out a good compromise among expressiveness, inferential capabilities and computability to use in the Web context.

The gap between software engineering methodologies based on the above languages and real-world ontology modeling is still a debated issue [17]. Web ontology representations have to deal with a spectrum of drawbacks spanning from language inherent ambiguity to context dependency, presence of incoherent statements, scattered pieces of information, difficulty in ontology matching and so on. It seems that all these issues have twofold reason: they lay both on the semantic (ontology) level and on the lexical (language) level. Consequently, semantic annotation of Web resources is prone to produce weak structure metadata. This is particularly true for collaborative (wiki) approaches [18] where personal conceptualizations are rather difficult to be mapped one another. Although such collaborative environments represent a challenge for the research community, they are still tailored to generic semantic services [19]. A top-down solution is to provide the tagging system with a well-defined and widely-accepted ontology: choosing the right ontology may be demanding in complex environment like the whole Web.

This work employs an agent-based architecture model for supporting the whole information retrieval process, from the user interface to the semantic tagging of Web resources. The agents that perform the annotation task use a Semantic Lexicon (hereinafter SL) as their golden ontology. In its actual implementation, the chosen SL was WordNet 3.0.

III. Using WordNet as golden ontology

The golden ontology paradigm focuses on comparing how well a given ontology resembles the gold standard in the arrangement of instances into concepts and the hierarchical arrangement of the concepts themselves [20]. A copious literature exists on golden ontologies [21][22][23]. In [24] Di Lecce and Calabrese address the new emerging approach of SL-based systems for modeling semantic Web applications. Starting from a preliminary survey on the different use of the concepts 'taxonomy' and 'ontology' in the literature, they identify SL as a good mediator between the two extremes. The authors also provide a SL-based abstract model suitable for multi agent system implementation. According to the authors' view, an indicative exemplar for the SL class is WordNet [2].

WordNet is a SL purposely engineered for text mining and information extraction. For example, it has been used to carry out Word Sense Disambiguation (WSD), for an overview of such characteristic the reader can refer to [25][26]. WordNet is referred to in the literature in several ways:

- Lexical Knowledge Base [27][28]
- Lexical Taxonomy [29][30]
- Lexical Database [31][32]
- Machine Readable Dictionary [33][34]
- Ontology [35][36]
- Semantic Lexicon [1][37]

Although, the above definitions can be considered synonyms, they emphasize different aspects of the same object. In this paper, only the latter definition will be used, since it accounts for the two elements (lexicon and semantics) which are relevant for the IR task, as explained forth. In this view, an important WordNet feature supplied by its underpinning data model is the capability of handling polysemy and synonymy. To this end, the concept of 'Sense Matrix' is introduced.

A. *Defining the Sense Matrix*

Two prominent causes of language ambiguity are polysemy and synonymy. Synonymy decreases recall and polysemy decreases precision, leading to poor overall retrieval performances [38]. It is interesting to note that synonymy represents a lexical relation among word forms while polysemy occurs when the same lexical form has multiple meanings. To define the relation among lexical and semantic entities at a finer grain, the definition of *Sense Matrix* is due. Thereby, a (*feasible*) *sense* is defined as particular element of such a matrix. Formally:

Def. (Sense Matrix). If *L represents the set of lexical entities and C the set of concepts of a given SL, a Sense Matrix S is defined as the matrix* $L \times C$ *such that* $S[i,j]=1$ *if* $(l_i, c_j) \in SL$ *and* $S[i,j]=0$ *otherwise. The set of feasible senses is defined as:*

$$FS := \{s_{ij} \mid S[i,j]=1\} \tag{3}$$

Throughout the text only feasible senses will be considered.

The concept of Sense Matrix is not new in the literature. In 2006 Swen [39] introduces almost the same notion. There is however some difference in terminology. The term 'sense' for Swen corresponds to our 'concept', thus, for Swen, a 'sense' is a term-document matrix. Our model can be considered as a specification to that of Swen assuming that senses are provided by a golden ontology.

It is noteworthy that *S* induces a binary matrix *M* on the Cartesian product *L* x *C* that is generally called 'lexical matrix' in the literature [40][41]. In [42], the lexical matrix is presented as an integral part of the human language system. Since there is no preference between the two dimensions represented by *M* (lexical and semantic), the authors prefer to refer to *M* as a Sense Matrix. This matrix can be considered as the base computational support for dictionary-based retrieval systems. Actually, it works as a look-up table that allows for switching from one dimension to another. An illustrative example of matrix *M* is provided in Table I.

TABLE I. EXAMPLE SENSE MATRIX . SENSES ARE DEFINED AS MATCHES BETWEEN LEXICAL ENTITIES (ROWS) AND CONCEPTS (COLUMNS)

SENSE MATRIX		CONCEPTS			
		c_1	c_2	c_3	c_4
LEXICON	l_1	0	0	0	1
	l_2	0	1	1	0
	l_3	1	1	0	0

B. *WordNet data model*

WordNet is organized around the idea of synsets, i.e. group of cognitive synonyms, each one representing a specific concept in a given context. Synsets are interlinked by means of conceptual-semantic and lexical relations. Any synset pertains to the concept layer, i.e. it is an instance of the set of concepts. The one-to-one relation between the synset and the word form produces the *sense*. Hence, a synset can be defined as the union of senses sharing the same concept entity (i.e. synonyms).

The 'sense' table combines tuples of the 'word' table with tuples of the 'synset' table. According to SL definition, the three tables define respectively the Sense Matrix, the set of lexical entities and the set of concept entities.

The WordNet taxonomic hierarchies (comprising the set of lexical and semantic relations) are covered by the two tables 'lexlinkref' and 'semlinkref'. Semlinkref defines only semantic relations, while lexlinkref defines lexico-semantic relations. In other words, lexlinkref provides recursive relations over the set of senses. An index to all kind of relations is contained in the 'linkdef' table.

WordNet is an ongoing project, since minor bugs and refinements characterize new version releases (in this work WordNet 3.0 was finally adopted). An excerpt of WordNet 3.0 class diagram is reported in Figure 1.

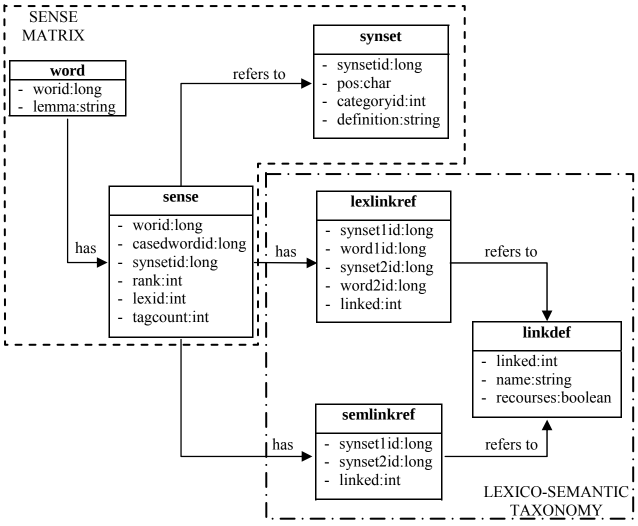


Figure 1. An extract of WordNet data model. Tables belonging to sense matrix and lexico-semantic taxonomy are grouped separately.

C. *Formal Knowledge Representation*

A formal representation of WordNet SL can be adapted from Dellschaft and Staab [21]. Referencing to a previous work on the subject [43], they provide a simple, but formal definition of a *Core Ontology* as the triplet:

$$CO = <C, root, \leq_C> \tag{4a}$$

where C represents the set of concepts, *root* the uppermost superordinate concept and $\leq_C$ a partial order on C (hence a taxonomical relation). Core Ontology seems to be an effective representation because it synthesizes the different layers constituting an ontology [44]. As a consequence of the introduction of the Sense Matrix, the Core Ontology definition used in this paper slightly changes. The set C is substituted by the set of feasible senses *FS*:

$$CO = <FS, root, \leq_C > \tag{4b}$$

It is evident that Core Ontology definition still has a taxonomical structure (actually a directed acyclic graph form)[1].

In WordNet, nouns, verbs, adjectives and adverbs can be considered as four lexical categories (also known as *part-of-speech* in the literature) each one defining a corresponding sub-ontology. Thus, the following partition generally[2] holds for a generic SL:

$$SL = \overline{O}_n \oplus \overline{O}_v \oplus \overline{O}_{adj} \oplus \overline{O}_{adv} \tag{5}$$

The overline is used to stress that the ontology is a golden one.

WordNet considers different semantic and lexical relation among concepts such as hyponymy/hypernym, meronym/holonym, antonyms, entailment and so on. Some relations are specific to certain categories like entailment for verbs; moreover, there are some relations having a single-rooted structure while some others are not. $\overline{O}_n$ is the only one having one single root (the synset conceptualizing the 'entity' lexical entry) hence, it suits the formal (4b) definition perfectly.

Notation. For the sake of conciseness, the following notation is introduced:

$$O^{rel}_{conc} \tag{6}$$

where *rel* represents a lexico-semantic relation, i.e. one element of the set {*hyponymy, hypermymy, holonymy, meronymy,...*}and *conc* is one of the four lexical categories, i.e. one element of the set {*nouns, verbs, adjectives, adverbs*}. O^{hyper}_n , for example, indicates an ontology defining hypernymy (relation) among nouns (concept nodes). In this paper only hypernymy has been employed in the considered golden ontology:

$$\overline{O}_n = \overline{O}^{hyper}_n \tag{7}$$

In [26] more WordNet relations are used with the aim of building *semantic graphs*. The authors adopt these structures in the Structural Semantic Interconnection (SSI) algorithm for the word sense disambiguation task. The semantic context is used in each iteration of the algorithm to disambiguate the lexical terms. Thus the accuracy of the algorithm is strictly related to the chosen context. The major difference between SSI and our approach is that, in the latter, the context is not an input for the sense disambiguation system.

IV. Proposed IR Model

The perfect search engines should respond to user query by listing exactly what the user actually queried for. Provided that this desirable situation is an ideal one, it is more feasible to reason about what current search engines *generally* do. They provide a ranked list of websites matching the user query according to a given algorithm. Upon search engines response, user chooses the website to browse, occasionally coming back to the search engine webpage to submit another query (Figure 2 reports an UML representation of the whole mechanism). Since the most of currently available search engines are not semantic-based, they index Web documents in a way similar to the Sense Matrix reported in Table I. User query is performed only at the lexical layer, being exposed to misinterpretation due to erroneous synonymy and polysemy interpretation. This means that the semantic gap is left totally to the user understanding.

[1] This complies with some relations like hypernymy/hyponymy and may be not sufficient for others where cyclic relations may occur. For the aim of this paper however, DAG structures only are considered. More complex grap-like structures are left to future work on the subject.

[2] It can happen that some relations (especially the morphological ones, like derivational forms) make this assumption not valid. In this sense, the provided partition should be intended as an opportune simplification for a working hypothesis.

In fact, in general a small text preview is fed back to the user, to let him/her decide the best option, basing on the semantics he/she gives to the displayed preview. This is an elegant way of bridging the semantic gap: the lack of this approach is that semantics is pushed in the query-response mechanism only from the user side. However, this “try and look” paradigm can be overcome in the light of the proposed IR system (Figure 3 depicts an UML representation of the mechanism characterizing the proposed IR system).

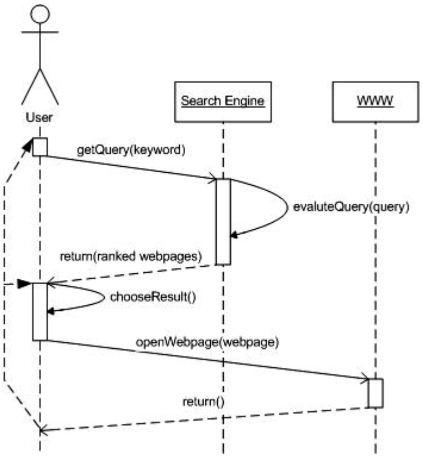


Figure 2. UML representation of the mechanism characterizing traditional search engines. The searching process starts with the user’s query. Keywords are the entry points for the search engine alghoritm.

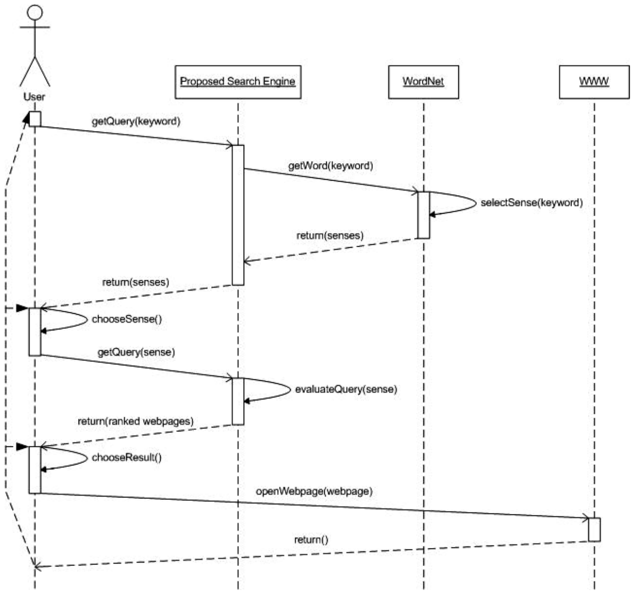


Figure 3. UML representation of mechanism characterizing the proposed semantic search engine. As in figure 2, the searching process starts with the user’s query. In this case, instead, senses related to the keywords are the entry points for the proposed semantic search engine alghoritm.

In the proposed system, as in traditional keyword-based systems, the user enters one or more keywords that he/she considers as significant for the kind of document he/she is searching for. Contrarily to retrieval systems based on term-document matrix, our IR system queries the golden ontology in order to get all possible senses related to (lexical) user query. The sense is explained by means of a short gloss, which is actually a meta-description of the sense itself. Once that the user has chosen the sense he/she wants to search for, the system retrieves all documents previously indexed by that sense.

A. New Browsing Paradigm

Our approach is based on a three dimensional dataset comprising respectively term, synset and document dimensions (Figure 4).

User query begins at the lexical level and then moves towards semantics thus becoming a two-step request/reply process:

1. The first step is a traditional keyword-based query performed at the lexical level. The system replies by listing the possible related senses.
2. The second step consists in user choosing the right sense thus entering the ‘semantic browsing mode’ which also allows for selecting the semantically indexed documents.

In this new framework, documents are indexed by senses. This does not affect the chosen document ranking criteria (like Page Rank) since dimensions are orthogonal. The real difference from traditional IR models is that user moves within a semantic space, eventually deciding to open a sense-related webpage (as examples related to a gloss in a dictionary).

System response in the first step is not possible unless some sense disambiguation technique is applied. For a previously published sense disambiguation technique, the readers may refer to the work of Di Lecce et al. [25].

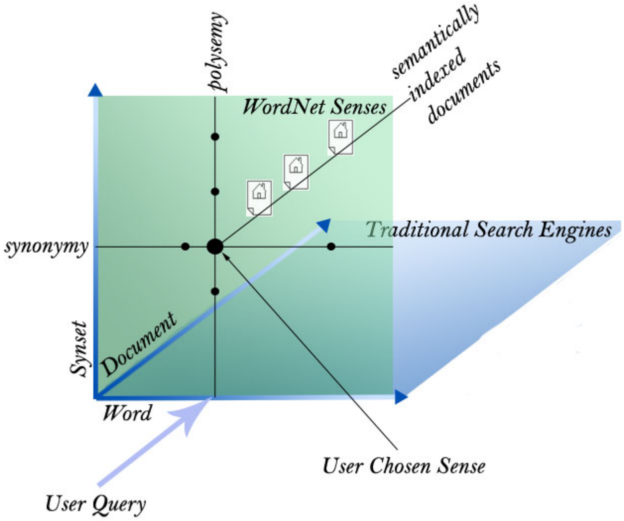


Figure 4. The proposed 3-dimensional IR model. It can be considered as the 3d space projection of the tuple *f(t, s, d)* characterized by term, synset and document.

V. SEMANTIC LEXICON-BASED MAS

MAS design has been gaining the attention of research community for many years [45]. Software agents are designed to cooperate (either with other agents or with humans) for managing the system knowledge base (KB) in different situations. In this paper a MAS implementation that employs WordNet as golden ontology is used to support the design of the proposed SL-based information retrieval system. The used MAS architecture is a hierarchical one [46] and is composed of the following layers:

- **Interface Layer**: it responds to user query. User may be human or computer such as crawlers and parsers;
- **Brokerage Layer**: it mediates among computational resources according to environment constraints;
- **Markup Layer**: it performs the tagging and other related activities.
- **Knowledge Layer**: it manages system knowledge base.

This agent-based approach is scalable because many features can be added to the SL-based system without affecting the underpinning model. For example, an inference engine may be added to the system in order to inference on new semantic relations among concept words. Tests in this direction are currently under way. Their aim is to assess the feasibility of domain-specific search engines that would enhance domain browsing and document retrieval.

The used MAS architecture has been inspired by previous works in other fields (see for example [46]). An overview of the proposed MAS architecture is depicted in Figure 5.

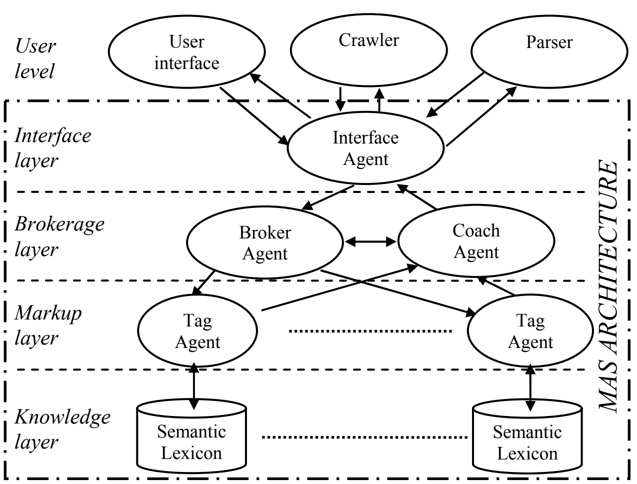


Figure 5. Multilayered MAS architecture used for semantically tagging Web resources. It is noteworthy the use of SL as golden ontology for the system knowledge.

Hereinafter an insight into the proposed MAS architecture is provided. Each different layer is described, with particular emphasis to the interface layer which handles user-system interaction, on the markup layer, which provides the Web resources semantic tagging process, and on the knowledge layer formally before presented.

A. User Level

It is the top level of the hierarchy and interacts with the MAS. It represents the communicational channel from and towards the system's environment. The user level is suitable composed of the three following elements: *user interface* – i.e. the human-machine interface; *spider* – a computer program that analyses the taxonomy structure of considered websites; *parser* – a computer program that extracts the relevant information from Web documents. The crawling/parsing processes are thoroughly described in [47]. Instead, a prototypal version of the user interface has been developed and presented in this paper (Figure 8 represents its actual implementation).

B. Interface Layer

The semantic browsing options handled by the interface agent are synthesized by the following Extended-BNF representation.

```
<interface> ::= <frame_header> <frame_www>
                <frame_semantics>
<frame_header> ::= {<sense>}
<frame_www> ::= {<href>}
<frame_semantics> ::=  [<forward_sense>]
                       [<backward_sense>]
<forward_sense> ::= {<sense>}
<backward_sense> ::= {<sense>}
<sense> ::= <word> <gloss>
```

The interface is composed of three frames:

1. Header: reporting the considered sense i.e. word-synset pair. The sense is described by means of the gloss associated to it;
2. WWW (traditional browsing): lists all Web resources indexed by the current sense;
3. Semantics (semantic browsing): allowing the user to move within the semantic space;

The choice of the extended version of the Backus Normal Form is due to the need to easily represent the cardinality for both elements sense and href. While curly brackets indicate the cardinality of a symbol, the square brackets represent the optional element in the derivation rule. *Forward_sense* and *backward_sense* are the parts of semantic space linked to the header sense. A graphical illustration explaining the BNF is presented in the next Section.

It is noteworthy that this interface shows recursive characteristics. The user can perform semantic browsing moving towards similar concepts in the query refinement process. In the experiment Section a screenshot of the prototypal implementation is commented in more detail.

C. Brokerage Layer

The MAS is triggered by user query submitted to the interface agent. Once the query has been correctly decoded, the interface agent leaves control to the Brokerage Layer. This layer is managed by two agents: broker and coach.

Broker Agent analyses which Tag Agents can satisfy the requirement. It manages all inbound communication coming from the Interface Agent. Starting by one query, it relays user service request to available resources of the lower layer, according to the chosen scheduling policy.

Coach Agent receives message from all the Tag Agents, collects and ranks the results. Next, it sends a message to the Broker Agent to inform that service request has been fulfilled.

Both agents of this layer are poorly detailed because that goes beyond the scope of this paper.

D. (Semantic) Markup Layer

In [25], a WSD algorithm was proposed to find the nearest common WordNet subsumer among words extracted from two link texts (also known as anchortexts). These couples of textual descriptions are taken from any possible pair of inbound and outbound links of a given webpage. If a concept subsumer (which is a synset) is found, lemmas, which lexicalize it, are used to tag the webpage.

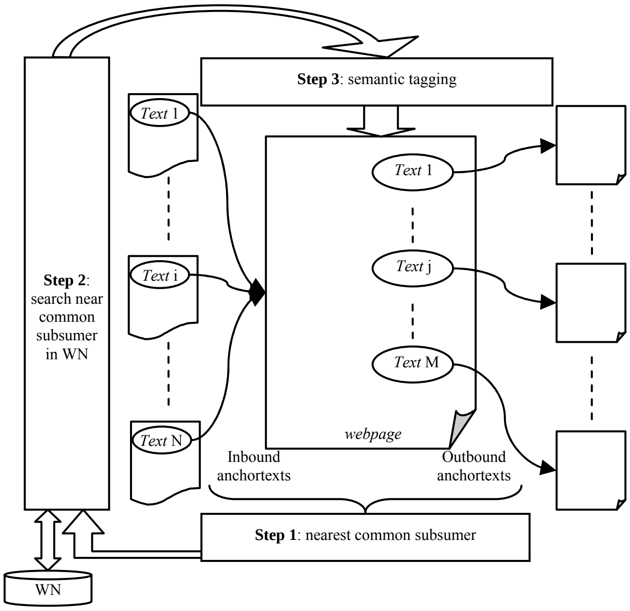


Figure 6. Actions performed by a Tag Agent. For any couple of inbound and outbound text links (step 1) the nearest common subsumer is searched in the WordNet database (step 2) according to the semantic relation pertaining the agent (e.g. hypernymy). If such synset element is found, its related lemmas are used to tag the corresponding webpage (step 3).

A *tag agent* (Figure 6 helps explaining how Tag Agent actually works) repeatedly performs this tagging activity on the list of webpages received by the *Broker Agent*.

For modularity purposes, each Tag Agent searches for semantic tags exploiting one of the possible semantic relations provided by the chosen SL. There can be one or more Tag Agents for hypernymy, others for holonymy and so on.

E. Knowledge Layer

Starting from the notations given in Section III.C, the system KB can be then formally expressed as follows:

$$KB = \overline{O}_n \bigcup_i \left\{ O_n^{href} \right\}_i \qquad (8)$$

$\left\{ O_n^{href} \right\}_i$ represents the individual ontology of the *i*-th inspected webpage; *n* represents the nouns (webpage semantic tags) identified by the tagging agent, finally *href* represents the HTML hyperlinks connecting tagged webpages. As shown in Figure 5 the Knowledge Layer can be split in many Semantic Lexicon units as much as Tag Agents exist in the upper layer. This ensures more flexibility and scalability of the system too.

VI. Experiments

Our experiments have been carried out on a data set of 48 distinct websites clustered in four semantic domains. Table II reports the number of analyzed websites for each semantic domain.

TABLE II. NUMBER OF INSPECTED WEBSITES FOR THE CHOSEN SEMANTIC DOMAINS

Semantic Domain	# inspected websites
University	17
Low-cost airline	10
Seaport	8
Airport	13

The crawling and the parsing phases have been limited to the analysis of the crawled first one hundred webpages for each website. This choice ensured the coverage of the main taxonomical structures of the inspected websites (general categories).

A. Prototypal Interface Implementation

According to the previously presented E-BNF representation a prototypal Web-based user interface has been implemented. Apache, PHP, MySQL and Ajax technologies have been used for this scope. Figure 8 shows an example screenshot of the user interface developed for the proposed semantic search engine. The interface can be divided into the following three frames:

1. *header frame*: it supports the user in the sense disambiguation process as specified in the UML of Figure 3. Thanks to this frame, the user selects the right sense in the synset-word (WordNet) plain depicted in Figure 4.

2. *www frame*: this area lists the hyperlinks indexed by the sense that results from the user query. By clicking one of the listed hyperlinks the user is redirected to the corresponding webpage as in traditional search engines (UML of Figure 2).
3. *semantics frame*: it allows the user for browsing the WordNet plain. The user is supposed to be in the sense chosen in header frame and can move forward or backward towards "neighbour" senses. Given two senses *a* and *b* they are considered here as neighbours if they exhibit a common subsumer. When the user selects a new sense the interface is reset (i.e. the user "is moved" to the new sense), thus showing a recursive behaviour.

B. *Semantic tagging experiments*

The semantic tagging process has been applied to the experiment set according to what was explained in the previous section. The considered semantic relations were hypernymy extracted from WordNet 3.0 release. The inspection depth in the WordNet taxonomy for finding the nearest common subsumer was thresholded to 5. This choice was affected by these reasons:

- Higher depth level in taxonomy accounts for very general concepts that are conceptually distant from the analyzed domain
- Computational effort may increase more than proportionally as depth level increases.

To evaluate the proposed architecture the presented results refer to the hypernym (*IS-A*) relation. Two different evaluations have been carried out during the test process. One of these is related to the evaluation of human agreement with the automatic markup system (qualitative test). The other one evaluates the amount of results given by the proposed system regarding the completeness of information (quantitative test).

Quantitative test. Figure 7 depicts the coverage index (in percentage) of the semantic markup grouped by the semantic domains defined in Table II. This value is useful to understand how much of the WordNet taxonomical structure is retrieved in the link-based architecture of a website. Coverage index (*ci*) has been computed for any webpage according to the simple formula:

$$ci\% = \frac{\#t}{\#w} \qquad (9)$$

where the numerator stands for the number of tagged webpages and the denominator equals the total number of inspected webpages for the website. Then, data have been grouped by domain. A box plot representation is adopted to have a synoptic view of mean, variance, minimum and maximum values for each domain. Moreover, it also represents outliers for the data set. It is noteworthy that nearly 25% of webpages are tagged by IS_A relations. As shown in Figure 7 the semantic domains of University, Seaport and Airport have a *ci* near to 28%, while semantic domain of low-cost airline has a lower *ci*. This is because inspected low-cost airline websites provide a flat cross-domain semantic structure (many heterogeneous services like car rental, hotel booking, tours, etc.).

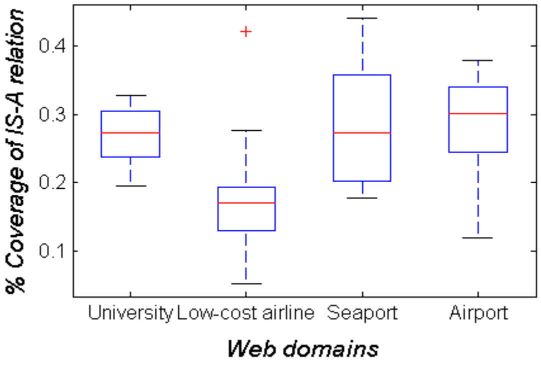


Figure 7. Coverage of IS-A relation for the considered semantic domains. The coverage values are considered in percentage. Any represented element is characterized by a continuous line (mean value) in a rectangle (variance range) and by two broken line ending with horizontal lines (minimum and maximum coverage for the data set).

Qualitative test. Table III reports an extract from the semantic mark-up process on the Manchester University website. Yet considering IS-A relation, the table is characterized by the uniform resource location (URL) of webpage, the lemmas associated to the sense markup and the anchortexts that caused the webpage to be tagged. The results are grouped by the URL identifier, in order to underline all sense markups assigned to a Web document. Table IV reports the semantic neighbors in WordNet semantic plain for data listed in Table III.

C. *Evaluation of the proposal*

The current semantic search panorama is quite a fragmented one. Although a lot of proposals can be found especially in recent years [49], they are still tailored to solving engineering aspects rather than being focused on performance. This conclusion can be fairly drawn by observing that even in the preface of the recent SemSearch 2008 International Workshop on semantic search [50] one of the major questions pointed out is: how can semantic search systems be evaluated and compared with standard IR systems?

Generally, small-sized comparison among the two approaches result in traditional IR systems largely outperforming semantic counterparts, at least for the recall performance. In [51] an attempt is made to fuse the two approaches by preserving the moderate recall of traditional system with the improved precision of the semantic-based ones. However, overall performances are still quite low and evaluation is confined to restricted datasets. In fact, one of the major pitfalls of corpus-based evaluation is the cost associated to the annotation. While in natural language processing (NLP) and word sense disambiguation (WSD) such datasets already exist and allow for good performances [52], in the general framework of Semantic Web it is currently impossible to think to a wide-covering semantic annotation for Web resources, at least until new standards like OWL will be sufficiently spread. In the meanwhile (which the authors think will last for many years ahead), the solution should be based on using available information at the maximum possible extent but from a different perspective, possibly by using new user-system interaction paradigms.

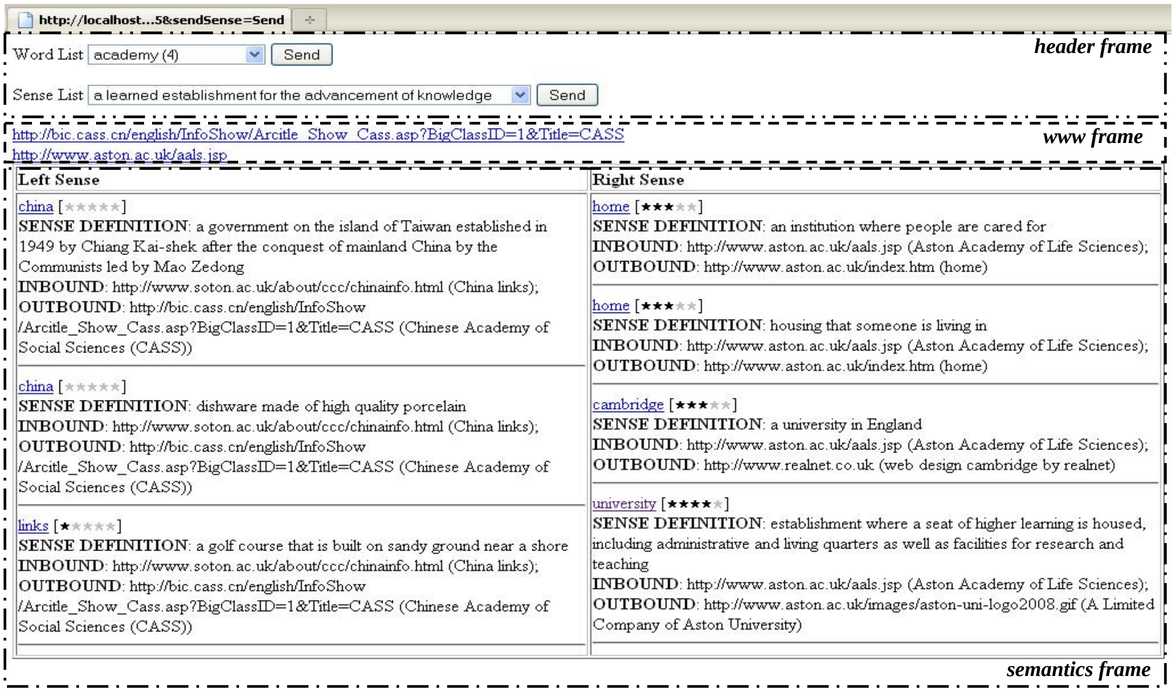
With reference to this paper, the focus was on the architecture that may leverage the simple mechanism of knowledge extraction and semantic annotation from the linked structure of the Web. This is an easy to run process which contributes to building a skeleton of semantic structures to which append (index) the crawled web pages. This allows the user to exploit a different navigation paradigm based on surfing a semantic graph rather than a web graph, thus reducing the semantic gap between the user and the retrieval system. Such an enhancement shifts the problem of retrieval to sense tagging and semantic disambiguation. More detailed numerical assessments of the proposed semantic tagging technique along with WSD aspects can be found in [47].

VII. CONCLUSION

In a recent survey on existing semantic search technologies [48], the authors categorize the 35 reviewed systems under three main facets: query, system, result. They conclude that a next step for the semantic search community is to foster the use of semantics in each of the three places. They also point out three main hinders to the evolution of semantic approaches: (1) lack of evaluation of semantic search algorithms, (2) lack of user evaluation of user interfaces, (3) lack of API and middleware support. The approach proposed in this work attempts to provide a way out to all these points by proposing a holistic framework centered on the idea of the SL-based architecture for Web IR. The main idea is to enlarge traditional TDM indexing structure up to a third dimension by adding a semantic layer. In this new model the user experiences a novel query paradigm which requires two consecutive steps: first to identify the sense related to documents he/she is querying for and then to access the semantically indexed document. In the line of a previous work specifically focused on semantic tagging of Web resources, this article proposes a MAS approach to Web IR design. Particular emphasis has been given to the interface layer managing user-system interaction and the markup layer performing the semantic tagging process. Since the proposed approach is highly modular, enlarging the experiment set will be the subject of our prospective research on this matter. The user interface will be also enriched and optimized in order to be effective for an extended number of inspected websites.

ACKNOWLEDGMENT

The authors thank Prof. Roberto Peluso for his useful remarks that helped us to better define several concepts during the writing of the paper. The authors would also like to thank the anonymous referees which allowed for improving, by their useful comments, the overall quality of the paper.



http://localhost...5&sendSense=Send

Word List academy (4)

Sense List a learned establishment for the advancement of knowledge

http://bic.cass.cn/english/InfoShow/Arcitle_Show_Cass.asp?BigClassID=1&Title=CASS
http://www.aston.ac.uk/aals.jsp

Left Sense	Right Sense
china [☆☆☆☆☆] **SENSE DEFINITION**: a government on the island of Taiwan established in 1949 by Chiang Kai-shek after the conquest of mainland China by the Communists led by Mao Zedong **INBOUND**: http://www.soton.ac.uk/about/ccc/chinainfo.html (China links); **OUTBOUND**: http://bic.cass.cn/english/InfoShow /Arcitle_Show_Cass.asp?BigClassID=1&Title=CASS (Chinese Academy of Social Sciences (CASS))	home [★★★☆☆] **SENSE DEFINITION**: an institution where people are cared for **INBOUND**: http://www.aston.ac.uk/aals.jsp (Aston Academy of Life Sciences); **OUTBOUND**: http://www.aston.ac.uk/index.htm (home)
china [☆☆☆☆☆] **SENSE DEFINITION**: dishware made of high quality porcelain **INBOUND**: http://www.soton.ac.uk/about/ccc/chinainfo.html (China links); **OUTBOUND**: http://bic.cass.cn/english/InfoShow /Arcitle_Show_Cass.asp?BigClassID=1&Title=CASS (Chinese Academy of Social Sciences (CASS))	home [★★★☆☆] **SENSE DEFINITION**: housing that someone is living in **INBOUND**: http://www.aston.ac.uk/aals.jsp (Aston Academy of Life Sciences); **OUTBOUND**: http://www.aston.ac.uk/index.htm (home)
links [★☆☆☆☆] **SENSE DEFINITION**: a golf course that is built on sandy ground near a shore **INBOUND**: http://www.soton.ac.uk/about/ccc/chinainfo.html (China links); **OUTBOUND**: http://bic.cass.cn/english/InfoShow /Arcitle_Show_Cass.asp?BigClassID=1&Title=CASS (Chinese Academy of Social Sciences (CASS))	cambridge [★★★☆☆] **SENSE DEFINITION**: a university in England **INBOUND**: http://www.aston.ac.uk/aals.jsp (Aston Academy of Life Sciences); **OUTBOUND**: http://www.realnet.co.uk (web design cambridge by realnet)
	university [★★★★☆] **SENSE DEFINITION**: establishment where a seat of higher learning is housed, including administrative and living quarters as well as facilities for research and teaching **INBOUND**: http://www.aston.ac.uk/aals.jsp (Aston Academy of Life Sciences); **OUTBOUND**: http://www.aston.ac.uk/images/aston-uni-logo2008.gif (A Limited Company of Aston University)

Figure 8. Screenshot of the prototypal interface. The three frames described in the text are confined in separate blocks. In the “semantics frame” each sense is quoted with a semantic relevance degree (star icons). Sense definition, along with original inbound and outbound anchortexts are also provided.

TABLE III. Excerpt from the semantic tagging process applied to Manchester University webpages on November 2008 [1]. The first column refers to the url of the tagged resources. The next column shows the found semantic tags. The 3rd and 4th columns report the inbound and outbound anchortexts respectively. In particular the lexical entries (words) that produced the semantic tag are capitalized and bolded. It is noteworthy that the same webpage may be referred to by more than a couple of anchortexts; hence it may be annotated by more sense tags.

Url	WordNet Sense Tag {# synset_id}	Anchortext 1	Anchortext 2
http://www.eps.manchester.ac.uk	ability, power {105616246}	computer **SCIENCE**	**FACULTY** of engineering and physical sciences
	bailiwick, discipline, field, field of study, study, subject, subject area, subject field {105996646}	computer **SCIENCE**	faculty of **ENGINEERING** and physical sciences
	body {107965085}	physics and astronomy **SCHOOL** of	**FACULTY** of engineering and physical sciences
http://www.langcent.manchester.ac.uk	body {107965085}	languages linguistics and cultures **SCHOOL** of	**UNIVERSITY** language centre
	construction, structure {104341686}		
	educational institution {108276342}		
	building, edifice {102913152}	languages linguistics and cultures **SCHOOL** of	university language **CENTRE**
	cognitive content, content, mental object {105809192}	find an academic department or **DISCIPLINE**	language **CENTRE**
http://www.manchester.ac.uk/aboutus/jobs/research	work {100575741}	**JOB** opportunities	**RESEARCH** jobs
http://www.manchester.ac.uk/aboutus/manchester/sport	activity {100407535}	**ART** and museums in manchester	**SPORT**
	activity {100407535}	nightlife and **ENTERTAINMENT**	**SPORT**
	diversion, recreation {100426928}		
	activity {100407535}	**NIGHTLIFE** and entertainment	**SPORT**
	diversion, recreation {100426928}		
http://www.manchester.ac.uk/aboutus/structure	artefact, artifact {100021939}	**ART** and museums in manchester	university **STRUCTURE**
	body {107965085}	**GOVERNANCE**	**UNIVERSITY** structure
	construction, structure {104341686}	**UNIVERSITY** structure	university **STRUCTURE**
	construction, structure {104341686}	university **STRUCTURE**	**UNIVERSITY** structure
	construction, structure {104341686}	**SUPPORT** services	**UNIVERSITY** structure
	construction, structure {104341686}	**SUPPORT** services	university **STRUCTURE**
	construction, structure {104341686}	chancellors of the **UNIVERSITY**	university **STRUCTURE**

TABLE IV. SUBSUMPTION HIERARCHY FOR NEIGHBOUR SENSES EXTRACTED FROM TABLE III. SENSE NEIGHBOURS ARE REPORTED IN THE LEFT COLUMN, WHILE THE RIGHT COLUMN ACCOUNTS FOR FIRST OR SECOND LEVEL COMMON SUBSUMER. SOME SENSE NEIGHBOURS SHARE THE SAME SYNSET SUBSUMER BOTH AT FIRST LEVEL AND SECOND LEVEL SYNSET DISTANCE.

Neighbour WordNet Sense	**WordNet Synset Common Subsumer**	
	First Level Distance	Second Level Distance
‘Science’, {105636887}	{105616246}	
‘Faculty’, {105650329}		
‘Science’, {105999797}	{105996646}	
‘Engineering’, {106125041}		
‘School’, {108275185}	{107965085}	
‘Faculty’, {108287586}		
‘School’, {108275185}	{107965085}	
‘University’, {108286163}		
‘School’, {102913152}	{104146050}	{104341686}
‘University’, {103297735}	{104511002}	
‘School’, {108277393}	{108276342}	
‘University’, {108286569}		
‘School’, {104146050}	{102913152}	
‘Centre’, {102993546}		
‘Discipline’, {105999266}	{105996646}	{105809192}
‘Centre’, {105921123}	{105809192}	
‘Job’, {100576717}	{100575741}	
‘Research’, {100633864}	{100636921}	{100575741}
‘Art’, {100908492}	{100933420}	{100407535}
‘Sport’, {100582388}	{100433216}	
‘Entertainment’, {100426928}	{100429048}	{100407535}
‘Sport’, {100582388}	{100433216}	
‘Entertainment’, {100429048}	{100426928}	
‘Sport’, {100523513}		
‘Nightlife’, {100426928}	{100582388}	{100407535}
‘Sport’, {100431292}	{100433216}	
‘Nightlife’, {100431292}	{100426928}	
‘Sport’, {100523513}		
‘Art’, {103129123}	{102743547}	{100021939}
‘Structure’, {104341686}	{100021939}	
‘Governance’, {108164585}	{107965085}	
‘University’, {108286163}		
‘University’, {103297735}	{104511002}	{104341686}
‘Structure’, {104341686}	{104341686}	
‘Support’, {104361095}	{104360501}	{104341686}
‘University’, {103297735}	{104511002}	
‘Support’, {104361095}	{104360501}	{104341686}
‘Structure’, {104341686}	{104341686}	

REFERENCES

[1] V. Di Lecce, M. Calabrese, and D. Soldo, "Semantic Lexicon-Based Multi-Agent System for Web Resources Markup", In Proceedings of the Fourth International Conference on Internet and Web Applications and Services (ICIW 2009), May 2009, Mestre, Italy (ISBN: 978-0-7695-3613-2), pp. 143-148.

[2] C. Fellbaum, WordNet: An electronic lexical database, MIT Press, Cambridge, (1998).

[3] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, (2008).

[4] H. P. Luhn, "The automatic creation of literature abstracts", IBM Journal of Research and Development, Vol. 2, No. 2, pp. 159-165, (1958).

[5] D. T. Tran, S. Bloehdorn, P. Cimiano, and P. Haase, "Expressive Resource Descriptions for Ontology-Based Information Retrieval", In Proceedings of the 1st International Conference on the Theory of Information Retrieval (ICTIR'07), October 2007, Budapest, Hungary, pp. 55-68.

[6] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", In Proceedings of the Seventh Internation World-Wide Web Conference (WWW 1998), April 1998, Brisbane, Australia, Computer Networks and ISDN Systems, Vol. 30, No. 1-7, pp. 107-117.

[7] A. Esuli and F. Sebastiani, "Page Ranking WordNet Synsets: An application to Opinion Mining", In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, June 2007, Prague, Czech Republic, pp. 424-431.

[8] S. Bloehdorn, M. Grobelnik, P. Mika, and D. T. Tran (editors), Preface to the Proceedings of the International Workshop on Semantic Search, located at the 5th European Semamntic Web Conference (ESWC 2008), June 2008, Tenerife, Spain.

[9] D. Tumer, M. A. Shah, and Y. Bitirim, "An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia", In Proceedings of the Fourth International Conference on Internet Monitoring and Protection (ICIMP 2009), May 2009, Mestre, Italy (ISBN: 978-1-4244-3839-6) , pp. 51-55.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, Vol. 41, No. 6, pp. 391-407, (1990).

[11] A. Graesser, A. Karnavat, V. Pomeroy, and K. Wiemer-Hasting, "Latent Semantic Analysis Captures Causal, Goal-oriented, and Taxonomic Structures", In Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society (CogSci 2000), August 2000, Philadelphia, PA, USA, pp. 184–189.

[12] R. Khare and A. Rifkin, "XML: a door to automated Web applications", in Internet Computing, IEEE, Vol. 1, Issue 4, July/August 1997, pp. 78-87.

[13] M. Vargas-Vera, et al., "MnM: Ontology-driven tool for semantic markup", In Handschuh, Mr Siegfried and Collier, Mr Niegel and Dieng, Miss Rose and Staab, Dr Steffen, Eds., Proceedings of the Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM 2002), Lyon, France, July 2002, pp. 43-47.

[14] K. Siorpaes and M. Hepp, "MyOntology: the marriage of ontology engineering and collective intelligence", Proceedings of the Workshop on Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007), at the 4th European Semantic Web Conference (ESWC 2007), Innsbruck, Austria, June 2007, pp. 127-138.

[15] S. Luke, L. Specter, and D. Rager, "Ontology-based knowledge discovery on the World Wide Web", In A. Franz & H. Kitano (Eds.), Working Notes of the Workshop on Internet-Based Information Systems at the 13th National Conference on Artificial Intelligence (AAAI96), AAAI Press, Portland, Oregon, August 1996, pp. 96-102.

[16] G. Antoniou and F. van Harmelen, "Web Ontology Language: OWL", in S. Staab and R. Studer, Handbook on Ontologies in Information Systems, Springer-Verlag, pp. 76-92, (2003).

[17] M. Diouf, K. Musumbu, and S. Maabout, "Methodological aspects of semantics enrichment in model driven", In Proceedings of the Third International Conference on Internet and Web Applications and Services, 2008 (ICIW '08), Athens, Greece, June 2008, pp. 205-210.

[18] R. Abbasi, S. Staab, and P. Cimiano, "Organizing resources on tagging systems using T-ORG", In Proceedings of the Workshop on Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007), at the 4th European Semantic Web Conference (ESWC 2007), Innsbruck, Austria, June 2007, pp 97-110.

[19] C. Lange, "Towards scientific collaboration in a semantic wiki", In Proceedings of the Workshop on Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007), at the 4th European Semantic Web Conference (ESWC 2007), Innsbruck, Austria, June 2007, pp. 119-126.

[20] J. Brank, D. Mladenic, and M. Grobelnik, "Gold standard based ontology evaluation using instance assignment", In Proceedings of the 4th Workshop on Evaluating Ontologies for the Web (EON2006), May 2006, Edinburg, Scotland.

[21] K. Dellschaft and S. Staab, "On how to perform a gold standard based evaluation of ontology learning", Proceedings of the 5th International Semantic Web Conference, (ISWC 2006) , November 2006, Athens, GA, USA, pp. 173-190.

[22] S. Farrar and D. T. Langendoen, "A linguistic ontology for the semantic Web", GLOT International Vol. 7, No. 3, March 2003, pp. 97-100.

[23] E. Zavitsanos, G. Paliouras, and G. A. Vouros, "A Distributional Approach to Evaluating Ontology Learning Methods Using a Gold Standard", 3rd Workshop on Ontology Learning and Population (OLP3), at the 18th European Conference on Artificial Intelligence (ECAI 2008), July, 2008, Patras, Greece.

[24] V. Di Lecce and M. Calabrese, "Taxonomies and ontologies in Web semantic applications: the new emerging semantic lexicon-based model", Proceedings of the IEEE International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC'08), December 2008, Vienna, Austria, (ISBN: 978-0-7695-3514-2), pp. 277-283.

[25] V. Di Lecce, M. Calabrese, and D. Soldo, "A Semantic Lexicon-based Approach for Sense Disambiguation and Its WWW Application", International Conference on Intelligent Computing (ICIC 2009), September 2009, Ulsan, Korea, pp. 468-477.

[26] R. Navigli and P. Velardi, "Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation", In IEEE Transactions on Patter Analysis and Machine Intelligence, Vol. 27, No. 7, July 2005, pp. 1075-1086.

[27] R. Basili, et al., "Knowledge-Based Multilingual Document Analysis", Proceedings of the International Conference On Computational Linguistics (COLING 2002) on SEMANET: building and using semantic networks - Volume 11, August 2002, Taipei, Taiwan, pp. 1-7.

[28] D. Inkpen, "Building A Lexical Knowledge-Base of Near-Synonym Differences", In Proceedings of the Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001), June 2001, Pittsburgh, PA, USA, pp. 47-52.

[29] J.J. Jiang and D.W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", In Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X), September 1997, Taipei, Taiwan, pp. 19-33.

[30] A. Gangemi, N. Guarino, A. Oltramari, and R. Oltramari, "Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level" In Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS-2001), October 2001, Ogunquit, Maine, USA, pp. 285-296.

[31] G. Miller, “WordNet: a lexical database for English.”, Communications of the ACM, Volume 38, Issue 11 , pp.39-41 (1995).

[32] N. Ordan and S. Wintner, “Representing Natural Gender in Multilingual Databases”, International Journal of Lexicography, Vol. 18, No. 3, pp. 357-370 (2005).

[33] J. Kegl, “Machine-readable dictionaries and education.” Walker, Donald E., Antonio Zampolli and Nicoletta Calzolari, eds., Automating the Lexicon: Research and Practice in a Multilingual Environment, Oxford University Press, New York, NY, USA, pp. 249 – 284 (1995).

[34] Y. Hayashi and T. Ishida, “A Dictionary Model for Unifying Machine Readable Dictionaries and Computational Concept Lexicons”, In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), May 2006, Genoa, Italy, pp.1-6.

[35] V. Snasel, P. Moravec, and J. Pokorny, “WordNet Ontology Based Model for Web Retrieval”, Proc. Of International Workshop on Challenges in Web Information Retrieval and Integration, (WIRI ‘05), April 2005, Tokyo, Japan, pp. 220-225.

[36] E. Nichols, F. Bond, and D. Flickinger, “Robust ontology acquisition from machine-readable dictionaries”, In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-2005), August 2005, Edinburgh, Scotland, pp. 1111–1116.

[37] T. Qian, B. Van Durme, and L. Schubert, “Building a Semantic Lexicon of English Nouns via Bootstrapping”, In Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium, June 2009, Boulder, CO, USA, pp. 37–42.

[38] G. L. Kowalski and M. T. Maybury, “Information Storage and Retrieval Systems. Theory and Implementation”, Springer, The Information Retrieval Series , Vol. 8, 2nd ed., (2000).

[39] B. Swen, “Sense Matrix Model and Discrete Cosine Transform”, In Proceedings of the first Asia Information Retrieval Symposium (AIRS 2004), October 2004, Beijing, China, LNCS 3411, Springer-Verlag, Berlin, Heidelberg, pp. 202-214.

[40] N. Ruimy, P. Bouillon, and B. Cartoni, “Inferring a Semantically Annotated Generative French Lexicon from an Italian Lexical Resource”, in Bouillon and Kanzaki (eds), Proceedings of the Third International Workshop on Generative Approaches to the Lexicon, May 2005, Geneva, Switzerland, pp. 27-35.

[41] B. Magnini, C. Strapparava, F. Ciravegna, and E. Pianta, A Project for the Construction of an Italian Lexical Knowledge Base in the Framework of WordNet, IRST Technical Report #9406-15, (1994).

[42] N. L. Komarovaa and M. A. Nowak, “The Evolutionary Dynamics of the Lexical Matrix”, Bulletin of Mathematical Biology, Vol. 63, No. 3, May 2001, pp. 451-485, Springer.

[43] A. Maedche and S. Staab, “Measuring similarity between ontologies”, In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web (EKAW ‘02), October 2002, Siguenza, Spain, pp. 251-263.

[44] J. Brank, M. Grobelnik, and D. Mladenić, “A survey of ontology evaluation techniques”, In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005), at 7th International Multi-conference on Information Society (IS’05), October 2005, Ljubljana, Slovenia, pp. 166-169.

[45] M. Wooldridge and N. R. Jennings, “Agent theories, architectures, and languages: a survey”, Intelligent Agents, Series Lecture Notes in Computer Science, Subseries Lecture Notes in Artificial Intelligence, Vol. 890, No. 8, Springer-Verlag, 1995, pp. 1-39.

[46] A. Amato, V. Di Lecce, C. Pasquale, and V. Piuri, “Web agents in an environmental monitoring system”, Proceedings of the International Symposium on Computational Intelligence for Measurement Systems and Applications (CIMSA 2005), July 2005, Taormina, Italy, pp. 262-265.

[47] V. Di Lecce, M. Calabrese, and D. Soldo, “Fingerprinting Lexical Contexts over the Web”, Journal of Universal Computer Science, vol. 15, no. 4 (2009), pp. 805-825.

[48] M. Hildebrand, J. van Ossenbruggen, and L. Hardman, “An analysis of search-based user interaction on the semantic Web”, Technical report. Information Systems. Centrum voor Wiskunde en Informatica (NL) (2007).

[49] M. Hildebrand, J. van Ossenbruggen and L. Hardman. “An Analysis of Search-based User Interaction on the Semantic Web”. Hildebrand, REPORT INS-E0706 MAY 2007. Centrum voor Wiskundeen Informatica Information Systems.

[50] S. Bloehdorn, M. Grobelnik, P. Mika, T. T. Duc, Preface of SemSearch 2008, CEUR Workshop Proceedings, online at CEUR-WS.org/Vol-334/

[51] F. Giunchiglia, U. Kharkevich, I. Zaihrayeu , “Concept Search: Semantics Enabled Syntactic Search”, SemSearch 2008, CEUR Workshop Proceedings, Vol-334, pp.109-123.

[52] R. Navigli, “Word Sense Disambiguation: a Survey”. ACM Computing Surveys, 41(2), ACM Press, 2009, pp. 1-69.