Aerospace Information System based on Semantic Technonogies and Ontology Management

A Web Portal for Semantic Search and Document Categorization

F. Gargiulo[1], G. Zazzaro[1], G. Romano[1], G. Gigante[1], A. Raggioli[2] and R. Fusco[2]
[1]*Soft Computing Lab, Italian Aerospace Research Centre – Via Maiorise, Capua (CE), Italy*
[2]*GruppoMeta, Via G. Porzio, 4 – Centro Direzionale, Napoli, Italy*

Keywords: Aerospace Lexical Domain Ontology, Word Sense Disambiguation, Aerospace Taxonomy, Semantic Search, Document Categorization, Aerospace Information System.

Abstract: This paper describes a semantic search tool based on our experience in using a new lexical domain ontology for aerospace integrated with an open source general purpose ontology to support aerospace engineers in the timely semantic retrieval of the knowledge. The semantic search module represents an integrated tool dedicated to the semantic search, extraction and classification of information and knowledge in aerospace domain. It describes the implementation of a disambiguation algorithm based upon these ontologies and a new interesting graphical user interface for semantic searches is presented. Furthermore, next to the domain ontology, a taxonomy for classifying aerospace documents is also proposed. The document classification algorithm that leverages the deep integration between the proposed lexical domain ontology and taxonomy is also described. Finally, some considerations about the usage of the semantic search module by the side of domain experts, semantic experts or common users are reported.

1 INTRODUCTION

The growing demands of developing complex information systems saving costs and guaranteeing reliability leads to the adoption of different paradigms facilitating knowledge sharing, interoperability, completeness, and reuse. This paper presents a new integrated tool for semantic search, extraction and classification of information and knowledge in aerospace domain. It is based on:

a) The proposal of a new lexical ontology and a new taxonomy for aerospace domain;
b) The integration of both of them with open source general purpose ontologies;
c) The implementation of a disambiguation algorithm based on these ontologies;
d) The implementation of a classification algorithm that leverages the deep integration between the new ontology and the new taxonomy;
e) A graphical user interface that allows natural language queries and "by meaning" queries on a very large number of documents (books, papers, news, websites, etc.) both for experts and common user.

The tool represents a subsystem of a wider architecture that was implemented in SIA portal. As described below, SIA – *Sistema Informativo Aerospaziale*, Aerospace Information System (sia.cira.it) – is software infrastructure for access, retrieval and exploitation of technical, scientific information for aerospace and high-tech user community and related domains and added value user services.

It comprises the following major subsystems:

I. Web Portal subsystem;
II. Semantic Search subsystem;
III. Linked Open Data subsystem;
IV. Document Warehouse subsystem.

This paper mainly focuses on the Semantic Search subsystem and provides a only brief description of the other subsystems.

Gargiulo F., Zazzaro G., Romano G., Gigante G., Raggioli A. and Fusco R..
Aerospace Information System based on Semantic Technonogies and Ontology Management - A Web Portal for Semantic Search and Document Categorization.
DOI: 10.5220/0004994703410348
In *Proceedings of 3rd International Conference on Data Management Technologies and Applications* (DATA-2014), pages 341-348
ISBN: 978-989-758-035-2

2 RELATED WORKS

The ontology term is borrowed from philosophy, where an Ontology is a systematic account of Existence. In Artificial Intelligence context we can describe the ontology of a program by defining a set of representational terms. In particular, an ontology is an explicit specification of a conceptualization of a domain of interest (Gruber, 1993).

An ontology can be written for different tasks since many domains may need a specific and formal representation of knowledge:

Data Integration: the purpose is to integrate heterogeneous information systems. Often different databases retain the same type of information in different patterns of data modeling. An ontology can be used as mediator between database schemas, allowing you to integrate information in different patterns and to realize an interpreter between data from two different sources.

Information Retrieval (IR): IR is the set of techniques used for the recovery information in electronic format. IR is the largest field of application of ontologies because they improve the accuracy of online searches by adding semantic information which is useful to reduce the search space.

Semantic Web: ontologies can be used to solve various problems of heterogeneity of the Web. Ontologies can enrich internal representation (metadata) of meaningful semantic labels, can build representations to model users with respect to their information needs and build mechanisms of mediation between metadata and information needs of the user (to build custom interfaces).

There are different types of ontologies depending on abstraction level (Guarino, 1998):

Top-level: ontologies with very general or abstract concepts such as space, time, behavior, action, etc. which are independent from specific domains, so as to be useful for their reusability in other ontologies. For this reason they are also called Meta-Ontology. Such ontologies alone may have little use but they are great for building knowledge bases.

Domain and Task Ontology: this type describes the vocabulary related to a generic domain (e.gg aerospace, medicine, geography) or a generic problem (e.g. diagnosis, configuration) and it can specify concepts of a top-level ontology.

Application Ontology: this kind of ontology describe concepts in a specific domain and the problems derived from it.

The use of ontologies to provide a single and shared representation of knowledge for all system components has been largely motivated in literature in the last decade: an interesting review of the state of art of ontology-based software engineering can be found in (Calero, 2005; Castañeda , 2010; Gasevic, 2009; Farfeleder, 2011), and the last recent proceedings of international forums like SWESE, W3C, SEKE discussing synergies between ontology engineering and software engineering. Synergies are discussed focusing on different key concerns.

The first concern is related to the development of life cycle integrating the adoption of ontologies.

The second one proposes methods to develop ontologies. Literature recognizes mainly two approaches: the experience-based (Gómez-Pérez, 2004) and the "engineered" based which defines a set of life cycle activities aiming at prototype refinement (Uschold, 1996; Noy, 2001).

The third concern is related to the development of ontologies. Literature proposes ontologies with different richness of expressivity and to different purposes. Lightweight ontologies are principally taxonomies, they include concepts, relationships between concepts, and properties describing concepts. Heavyweight ontologies are those which model knowledge and define restrictions on domain semantics, by means of axioms and constraints. Ontologies are developed to support the development process, to support the knowledge sharing of general information about "the real world" and the application domain (medicine, automotive, railway, aerospace) (Calero, 2005).

The fourth concern aims to develop a complete framework proposing both new methodologies and tools to guide the use of ontologies and to apply it to each phase of software life cycle (Gasevic, 2009). In the aerospace industry domain ontologies are a constant in each approach but are rarely defined. An important work describing a basic ontology for aerospace is presented in (Malin, 2006) where the basic concepts of functions, entities and problems are defined. Specific ontologies are proposed to support the justification of design, RaDEX (Kuofie, 2010), and the aerospace composite manufacturing domain (Verhagen, 2011), to define UAV missions (Schumann, 2012) and to support the intelligence, surveillance, and reconnaissance (ISR) mission.

NASA addresses the use of ontologies in different contexts. The CDXA program aims to integrate knowledge in complex programs proposing a constellation of ontologies (SWEET, 2011).

Recent European projects adopt sematic techniques in aerospace domain. The EU CESAR project bring innovations in the two most improvable engineering disciplines: Requirements engineering and Component-based design of automotive, aerospace and railway (Bogusch, 2011).

3 DESCRIPTION OF THE SIA PROJECT

SIA project activity has been conceived as a strategic information tool in order to facilitate the growth of aerospace knowledge in the Regione Campania as it is oriented to the main aerospace actor such as SME, Universities, Research Centers, etc. This work has been carried on within the research project SIA, funded by the Campania Region and EU within the framework of POR Campania FESR 2007 – 2013.

SIA is SW infrastructure for access, retrieval and exploitation of technical, scientific information for aerospace and high-tech user community and related domains and for providing them added value user services.

The objective of SIA project is the development of a SW infrastructure able:

- to guarantee access to the most important source of information related to the aerospace domain (paper, technical report, e-books, e-journals);
- to facilitate the exchange among users of knowledge related to the research and development activities in the aerospace domain;
- to support user in the optimization of complex activities such as certification task, e-learning, etc.

SIA services are accessible through a vertical web portal based on semantic features with aerospace ontology and taxonomies with which SIA system:

- make documents content more meaningful for an efficient search and access of information;
- provide to users a relevant information as result of a search avoiding the negative experience of information overload or out of scope;
- enables users to activate strategies for a more efficient sharing and spread of domain knowledge.

Furthermore, SIA web portal offer to end-users the following functionalities:

1. user profile management for adaptive access to SIA services and content;
2. advanced information retrieval able to more exploit semantic information representative both of documents contents managed by SIA and user preferences;
3. semantic enterprise wiki in order to facilitate information exchange among users and to improve collaborative work;
4. Press and automatic News generation and aggregation for a more wide information spread.

In regard to 2 and 4 above, the user can define an alert (i.e. a set of semantic queries) and she will be notified when the system will index any document that satisfies the defined alert.

SIA operational context is built upon the following software subsystems conceived to satisfy project requirements:

Web Portal Subsystem: SIA portal through which services are made available to users (Search, Browsing, Blog, Wiki, News, News Alert, e-Press, Reference). Access to SIA web contents and services depends on user profile regardless of user device (PC, PDA, Tablet, etc.).

Semantic Search Subsystem: this component guarantees features in terms of automatic retrieval of predefined information sources, content filtering, parsing, word disambiguation, data extraction and correlation, data classification, indexing, data storage.

Linked Open Data Subsystem: a triple store based on Virtuoso and a SPARQ endpoint for sharing information about the document indexed by Semantic Search subsystem in RDF format according to W3C best practice regarding semantic interoperability.

Document Warehouse Subsystem: assures loading and storage of structured information generated in the Semantic Search Subsystem in a document warehouse for further user analysis tasks based on OLAP features.

4 THE SEMANTIC SEARCH SUBSYSTEM

The Semantic Search subsystem, as mentioned earlier, is characterized by the integration between the proposed lexical ontology and open source

general purpose ontologies joined to the disambiguation and classification algorithms and the graphical user interface. The following paragraphs describe the features of each of these aspects.

4.1 Aerospace Lexical Domain Ontology

The ontology development was inspired by the steps identified by Pinto & Martins (Pinto and Martins, 2004), which is roughly reflected in this section.

Specification: the aerospace domain ontology has a twofold objective. From a broad perspective, the purpose of this ontology is to fill a gap by introducing a new domain ontology. In a more strict sense, the purpose of the ontology is to support knowledge management, allowing the indexing, disambiguation, classification and search for contents in aerospace domain. The scope of the ontology is limited to its domain, and within this scope, the emphasis lies on the concepts and relationships of meaning among them.

Conceptualization: the applicable concepts and relationships come from an amalgam of various sources. The domain experts involved into the SIA project taken very carefully into account the existing ontologies (SWEET, 2010; Hannessen, 2003; CIRA, 2012). These ontologies have been studied in order to capture the current state of the art.

Formalization: the lexical domain ontology is available in two languages: Italian and English language. The following semantic relations in both languages are handled:

- Synonymy: it indicates the relationship between two terms that have the same or nearly the same meaning in aerospace domain, such as “space” and “cosmo”;
- Hypernymy: the relation of being superordinate or belonging to a higher (more abstract) rank or class. Inverse of hyponym. For instance, “tree” is hypernym of “oak” and “poplar”;
- Hyponymy: used to designate a member of a class. For instance, “Boeing 747” and “Airbus A380” are hyponym of “Aircraft”;
- Meronym: a word that denotes a constituent part or a member of something. For example, “wings” and “engine” are a meronyms of “Aircraft”;
- Holonym: the opposite of a meronym is a holonym, the name of the whole of which the meronym is a part.

Implementation: the domain ontology contains 7,497 Italian words, 5,962 Italian synsets, 5,750 English words and 5,127 English synsets. The Italian and English version of the ontology share 4,344 multilingual relations. In addition, the Italian version contains other 1,405 relations. The general purpose multilingual ontologies used are actually: the WordNet developed at the Princeton University (Fellbaum, 1999) for English language and the MultiWordNet for Italian language. For this reason, a natural choice was store the domain lexical ontology in the same database schema of MultiWordNet and therefore a custom simple ontology editor was developed. This editor allows to manage concepts, synsets and relations in the MySQL database schema. Also, the editor allows the editing of the domain taxonomy, described below, contained in the same database. A simpler version of the editor has been included in the Web Portal subsystem and it is available for administrative tasks.

4.2 The Ontologies Integration

The lexical ontologies in the Semantic Search subsystem are three: WordNet for English language, MultiWordNet for Italian language and the aerospace domain lexical ontology for both languages. The WordNet contains about 117,000 synsets and the currently available release of MultiWordNet includes information about 58,000 Italian word meanings and 32,700 synsets.

During the morphological analysis, the disambiguation and indexing phase of an Italian document, the tool can rely on the MultiWordNet and the domain ontology to detect the concepts and assign the meaning to the words in the document; it uses the WordNet and the domain ontology for English language instead. Before the disambiguation phase, concepts that are compound words, abbreviations or acronyms are detected. These concepts are usually relevant in domain terminology and their identification permits a more accurate tokenization of the sentences. If a token is related to a concepts belonging both general purpose and domain ontology, the system performs a sort of context analysis to determine if a general or a domain specific meaning would to be assigned to the word. The system tries to guess if the context is strictly related to the domain analyzing the previous and the following sentences and counting the number of tokens related to the domain terminology. If this number exceeds a fixed threshold then the

system chooses the domain meaning otherwise the general meaning.

Even if the domain meaning was chosen, the system also memorize the general meaning; this information can be used during the disambiguation of the adjacent token that do not belong to the domain ontology.

At the end, the disambiguation algorithm assigns an identifier to the token. Note that some token does not have an identifier, in particular: the terms that were not recognized, articles, conjunctions, prepositions, etc. The identifiers follow the MultiWordNet pattern and are formed by part of speech followed by (#) character and a numeric string consisting of 8 characters, e.g. `n#00001234` (`"n" means "noun"`). Identifiers of domain ontology to be unique in the overall system are prefixed by "d" (domain) and a number. For example `d1n#00001234` represents a domain concept. The number allows the simultaneous presence of more domain ontologies and distinguishes which of domain ontologies the concept belongs (in this work there is only one domain ontology).

At this point, the elaboration of a natural language queries can be briefly explained. In fact when the user inserts a sentence and executes a natural language query, the sentence will be processed in the same manner described above and at the end of the elaboration the system knows which concepts – and their ontologies - must search for. In other words, the system tries to determine the correct meaning of each word of the sentence from the sentence itself and therefore the sentence represents the context wherein the disambiguation is performed. On the other hand, if the user inserts a word and executes a "by meaning" query the system prompts the user to select one or more meanings among those present in ontologies, as described in more detail later.

4.3 The Disambiguation Algorithm

The disambiguation algorithm adopted is an implementation of a variant of the JIGSAW algorithm for word sense disambiguation proposed by University of Bari (Basile, 2007). For reasons of space the algorithm will not be described here (please, for the details refer to the original paper and to the documentation of the University of Bari). In this paper only the main changes occurred during the implementation will be described. The changes were aimed at improving integration with the other components of the system and an improvement of the performance.

The first adjustment involved the modules assigned to morpho-syntactical analysis and tokenization. The original system had its own morpho-syntactic analysis module, this module has been removed and replaced by two modules, respectively: Gate for English language and TreeTagger (Schmid, 1995) for Italian language.

With regard to the performance, it was noted that the introduction of a caching mechanism runs the algorithm about 5 times faster. This mechanism helps to avoid re-running the analysis of a token if a syntactic constructs (with equivalent terminology) was analyzed above. Therefore, could be formulated a conjecture about the high frequency of re-use of terminology in very specialized domains documents (such as aerospace).

4.4 The Classification Algorithm

A Bayesian model is trained in order to classify the indexed documents in taxonomy categories, Table 1. It is based on the Weka (Hall, 2009) implementation of the Bayesian multinomial classifier named NaiveBayesMultinomial (McCallum , 1998).

Table 1: The domain taxonomy.

Level 1	Level 2	Level 3
Aerospace		
	Aeronautics	
		Physics
		Materials
		Propulsion
		Equipment
		Design and Validation
		Traffic Management & Airports
	Space	
		Physics
		Materials
		Propulsion
		Equipment
		Design and Validation
		Ground Support & Launch Operations
	Sciences	

During the training phase of the classification model a standard training set based on an association between documents and classification taxonomy categories was not used. In fact such a kind of training set requires a huge number of documents manually tagged with the category. A different approach instead was proposed; it is based on the

presence of domain ontology and semantic disambiguation system. As mentioned, the disambiguation algorithm determines if a word can be associate to a concept belonging to the domain and this information can provide a significant contribution to the attribution of a document to a specific category. In fact, in the domain terminology are often present terms closely related to some categories, such as the name of the specific missile propellant, on-board equipment, etc. The presence of such terms often accurately directs the document to a category and, at the same time, filters out potentially noise resulting from the generic terminology.

Then, the domain ontology concepts were associated with the taxonomy categories with a weight that represents the degree of membership. In a number of cases it was not possible to create this association because the concept is too general (i.e., "Flight") or the domain experts did not found the association. About 2,880 domain ontology concepts are associated to one or more taxonomy categories and with these concepts the training sets – one for each category – are built. In particular, the training set of a category contains the concepts associated to that category and the weight of the association represents the label. When a document has to be classified, the system detects all domain concepts associated to one or more categories and submits them to the classification model. Finally, the classifier evaluates the degree of membership of the document to every taxonomy category.

4.5 The Semantic Search

Semantic Search function is the core of SIA. User can search documents through input query written in natural language or keyword-based. Until now, about 800,000 documents in both languages were indexed and it has been designed in order to aid user in the search of useful information and documents. At this end, in the SIA system three search features have been implemented:

Natural Language Search: an user can search documents through input query written in natural language. This search activates semantic disambiguation of the user input text and the system founds documents containing semantically disambiguated terms consistent to the context analysis performed on input text.

Lemma Search: SIA identifies the different meanings of each user input text through a querying into a general ontology (WordNet and MultiWordNet) and a domain ontology (SIA aerospace ontology). SIA shows an interactive window with all possible meaning and relations related to the search term. User can select a specific meaning (lemma) and its semantic relationship with other lemma in order to refine search. In SIA, this kind of query is also called “by meaning”.

Keywords Search: traditional full-text search performed on the basis of user query terms.

Whatever is the search feature selected by the user, SIA returns search results grouped by predefined facet: data source, format (html, pdf, word, etc.), category (main aeronautical and space taxonomy class), type (magazine, journal, etc.), authors, keywords, domain entity. Furthermore, search results are ordered according to a score function evaluated on the basis of the Virtuoso scoring algorithm. After the user selects a document, the system redirects her to a detail page. In this page there are also a list of similar document to the selected one. As described previously, the Natural Language Search lead back to Lemma Search therefore the latter will be described in more details.

The user executes the following steps: selects this kind of search, types a word and chooses the language. At this point, the system will guide her in the choice of the meaning or meanings of the word that should be looking for. It displays a tree like the one shown in the figure 1.

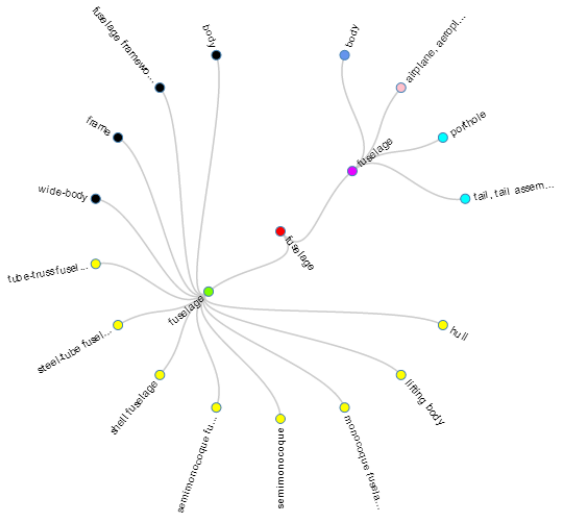

Figure 1: The GUI of lemma search in SIA.

All nodes of this tree, but the root, are concepts contained in the general or domain ontologies.

On mouse move the system shows a tooltip for each node with the exact definition of the concept and the lemmas it contains, figure 2.

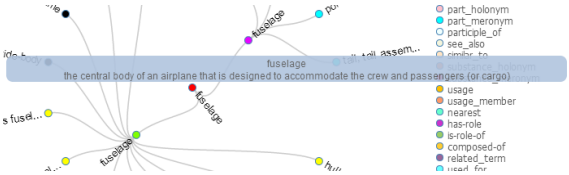

Figure 2: The tooltip with definition of the general meaning of the term (the pink top node) “fuselage”.

The red node is the root of the tree and it represents the term T that the user entered (“fuselage” in figure 1). The children of the root represents all the broad meanings of the term T. In practice, for each child C of the root the term T is a lemma of C (i.e., T belongs to the synset of C). Moreover, if D is a child of C it means that there is a relationship between C and D (in one of the two ontologies) and the color of D represents the type of relationship. As mentioned, the semantic relations handled are: synonymy, hypernymy, hyponymy, holonymy, meronymy; it is possible to think the concept D as a "specialization" of the concept C (for example, D is "part of" C or D is an hyponym of C).

Note that the term T can belong to both concepts C and D. In this case, the root is connected only to the node C because C is a more general concept of D.

Each node in the tree, but the root, can be selected and the search will return documents in which the term T is used in the meanings that correspond to nodes/concepts selected.

The interpretation of the edge between the root and its children is different with respect to the interpretation of the edges the links a child of the root to its children. In the first case, the children of the root are the broad meanings of the term T. At this level the user manually disambiguate the term and the color of the child indicates the respective ontology, whereby the user can decide whether to continue the search on only one of the ontologies, or on both.

The other edges always represent semantic relations and the color of the children of a child of the root specifies the kind of relationship. At this level, the user refine the meaning of the concept that must be searched for.

This type of representation makes it possible to distinguish the concepts that belong to the domain ontology from those that belong to the MultiWordNet. In fact, the concepts of domain ontology have different colors are drawn in a different side of the screen than the general ontology, figure 1. The layout used to draw the tree will tend always to separate the nodes of an ontology from those of the other. The number of nodes in the domain ontology compared to the number of nodes of the general ontology provides intuitively a measure of how much the current search is relevant with the aerospace domain.

This graphical user interface can be used by expert users who are familiar with the use of semantic relations; by user who are familiar with domain terminology; by common users.

For the first group of users, the GUI lets them select the semantic relations while for the other groups of users, they simply selects the meanings of a term and can freely ignore relations. In any case, it is clear that the lower nodes of the tree the more restrictive and specific meanings is associated to the term.

5 CONCLUSIONS AND FUTURE WORKS

This paper describes an integrated SW tool aimed to semantic search, extraction and classification of information in the aerospace domain. A new lexical aerospace domain ontology is proposed. The tool is based on the integration between lexical ontologies and algorithms operating on them and, in order to get a better user experience, a GUI is also proposed.

Experimental results regarding the performances of the tool was obtained in two ways. Precision and Recall measure (Davis, 2006) were firstly calculated on a test set consisting of 50 documents. The results were approximately the same in (Davis, 2006). Due to the small number of documents, these measures were not been interpreted as a measure of the performance of the disambiguation algorithm but rather as a confirmation that the changes introduced (i.e., Gate, TreeTager and caching) did not worsen the original algorithm. On the other hand, the response of domain users who have followed the experimental phase, represented a very positive qualitative assessment of the tool. In order to obtain consistent experimental results about the global performance of the described semantic search module it is necessary a comparison between similar systems developed for the aerospace domain that expose similar functionalities (lemma search, natural language search, etc.). Also, the employed set of ontologies plays a central role in the performance of the each system and how effective is a comparison among systems based on different ontologies is not a trivial matter. On the other hand, for the disambiguation algorithm used in semantic search

module it is possible to refers to the experimental results provided in (Basile, 2007). In particular, the disambiguation algorithm has been evaluated by SemEval-2007 task. The algorithms were scored according to standard IR/CLIR measures as implemented in the TREC evaluation package (http://trec.nist.gov/).Future works will aim to the maintenance of the lexical ontology. Updating and correcting the implemented ontology will be achieved by the publication of the ontology editing functions and the preparation of a controlled change management process for the approval of changes suggested by users.

REFERENCES

Basile, P., de Gemmis, M., Gentile, A. L., Lops, P., Semeraro, G., 2007. *UNIBA: JIGSAW algorithm for word sense disambiguation*. Proceedings of the 4th International Workshop on Semantic Evaluations (pp. 398-401). Association for Computational Linguistics.

Bogusch, R., Gerlach, S., 2011. *Optimierungen im Requirements-Engineering in der Praxis*.

Calero, C., Ruiz, F., Piattini, M., 2005. *Ontologies in Software Engineering and Software Technology*. Springer.

Castañeda, V., Ballejos, L., Caliusco, M., Galli, M., 2010. *The Use of Ontologies in Requirements Engineering*. Global Journal of Researches in Engineering, Vol. 10 Issue 6.

CIRA, 2012. *Thesaurus: descrittori logici delle attività aerospaziali.* Italian Research Aerospace Centre internal technical document.

Davis J., Goadrich M., 2006. *The Relationship Between Precision-Recall and ROC Curves*. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA.

Farfeleder, S., Moser, T., Krall, A., 2011. *Ontology-Driven Guidance for Requirements Elicitation*. ESWC, Part II, LNCS 6644, pp. 212–226 Springer-Verlag Berlin Heidelberg.

Fellbaum, C., 1999. *WordNet*. Blackwell Publishing Ltd.

Gasevic, D., Kaviani, N., Milanovi, M., 2009. *Ontologies and Software Engineering*. International Handbooks on Information Systems, pp 593-615 Springer.

Gómez-Pérez, A., Fernández-López, M., Corcho, O., 2004. *Ontological Engineering*. Springer, New York. USA.

Gruber, T. R., 1993. *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. International Journal Human-Computer Studies 43, p.907-928.

Guarino N., 1998. *Formal Ontology and Information Systems*. Proceedings of FOIS'98, Trento, Italy. Amsterdam, IOS Press, pp. 3-15 11.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. *The WEKA data mining software: an update*. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Hannessen, D. P., Donker, J. C., 2003. *ACARE Taxonomy A common European taxonomy for aeronautical research & technology*.

Kuofie, E. J., 2010. *RaDEX: A Rationale-based Ontology for Aerospace Design Explanation*. Master of Science Programme Business Information Technology University of Twente. http://essay.utwente.nl/59926/

Malin, J., Throop, D., 2006. *Basic Concepts and Distinctions for an Aerospace Ontology of Functions, Entities and Problems*. Aerospace Conference, IEEE.

McCallum, A., Nigam, K., 1998. *A Comparison of Event Models for Naive Bayes Text Classification*. AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).

Noy, N., McGuinness, D., 2001. *Ontology development 101: A guide to creating your first ontology*.

Pinto H. S., Martins J. P., 2004. *Ontologies: How can they be built?* Knowledge and Information Systems, Vol. 6, No. 4, pp. 441–464.

Schmid, H., 1995. *TreeTagger| a Language Independent Part-of-speech*. Tagger.Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43.

Schumann, B., Scanlany, J., Fangohrz, H., 2012. *A Generic Unifying Ontology for Civil Unmanned Aerial*. 12th AIAA Aviation Technology.

SWEET Ontologies - Nasa, 2011. URL: sweet.jpl.nasa.gov/ontology.

Uschold., M., 1996. *Building ontologies: towards a unified methodology*. Proceedings of 16th Annual Conference of the British Computer Society Specialists Group on Expert Systems.

Verhagen, W. J., Curran, R., 2011. *Ontological modelling of the aerospace composite manufacturing domain*. Improving Complex Systems Today (pp. 215-222). Springer London.

Bloehdorn, S., Hotho, A., 2006. *Boosting for text classification with semantic features*. Advances in Web mining and Web usage Analysis (pp. 149-166). Springer Berlin Heidelberg.

Song, M. H., Lim, S. Y., Kang, D. J., Lee, S. J., 2005. *Automatic classification of web pages based on the concept of domain ontology*. Software Engineering Conference, 2005. APSEC'05. 12th Asia-Pacific (pp. 7-pp). IEEE.

Wu, S. H., Tsai, T. H., Hsu, W. L., 2003. *Text categorization using automatically acquired domain ontology*. Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11 (pp. 138-145). Association for Computational Linguistics.