www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

KNOWLEDGE REPRESENTATIONS WITH ONTOLOGY SUPPORT FOR COLLABORATIVE ENGINEERING IN ARCHITECTURE ENGINEERING AND CONSTRUCTION

PUBLISHED: September 2014 at http://www.itcon.org/2014/26
EDITOR: Rezgui Y.

Ruben Costa, PhD student
Centre of Technology and Systems UNINOVA, Portugal
rddc@uninova.pt

Celson Lima, Professor
Federal University of Western Pará UFOPA / IEG / PC, Brasil
celson.lima@ufopa.edu.br

SUMMARY: *This paper brings a contribution focused on collaborative engineering projects where knowledge plays a key role in the process. Collaboration is the arena, engineering projects are the target, knowledge is the currency used to provide harmony into the arena since it can potentially support innovation and, hence, a successful collaboration. The Building and Construction domain is challenged with significant problems for exchanging, sharing and integrating information among actors. Semantic gaps or lack of meaning definition at the conceptual and technical level, for example, are problems fundamentally originated through the employment of representations to map the 'world' into models in an endeavour to anticipate other actors' views, vocabulary, and even motivations. One of the primary research challenges addressed in this work relates to the process of formalization and representation of document contents, where most existing approaches are limited and only take into account of the explicit, word-based information in the document. The research described in this paper explores how traditional knowledge representations can be enriched through incorporation of implicit information derived from the complex relationships (Semantic Associations) modelled by domain ontologies with the addition of information presented in documents, by providing a baseline for facilitating knowledge interpretation and sharing between humans and machines. The paper introduces a novel conceptual framework for representation of knowledge sources, where each knowledge source is semantically represented (within its domain of use) by a Semantic Vector. This work contributes to the enrichment of Semantic Vectors, using the classical vector space model approach extended with ontological support, employing ontology concepts and their relations in the enrichment process. The test bed for the assessment of the approach is the Building and Construction, using an appropriate domain Ontology. Preliminary results were collected using a clustering algorithm for document classification, which indicates that the proposed approach does improve the precision and recall of classifications. Future work and open issues are also discussed.*

KEYWORDS: *Construction Industry, Knowledge Sharing, Semantic Interoperability, Ontology Engineering, Unsupervised Document Classification, Vector Space Model*

REFERENCE: *Ruben Costa, Celson Lima (2014). Knowledge representations with ontology support for collaborative engineering in architecture engineering and construction, Journal of Information Technology in Construction (ITcon), Vol. 19, pg. 434-461, http://www.itcon.org/2014/26*

1. INTRODUCTION

Over the last two decades, the adoption of the Internet as the primary communication channel for business purposes brought new requirements especially considering the collaboration centred on engineering projects. By their very nature, such projects normally demand a good level of innovation since they tackle highly complex challenges and issues. On one hand, innovation often recurs to combination of knowledge (existing, recycled, or brand new) and, on the other hand, it depends on individuals (or groups) holding the appropriate knowledge to provide the required breakthrough.

Engineering companies are project oriented and successful projects are their way to keep market share as well as to conquer new ones. Engineering projects strongly rely on innovative factors (processes and ideas) in order to be successful. From the organisation point of view, knowledge goes through a spiral cycle, as presented by Nonaka and Takeuchi (Nonaka & Takeuchi, 1995). It is created and nurtured in a continuous flow of conversion, sharing, combination, and dissemination, where all the aspects and contexts of a given organisation, are considered, such as individuals, communities, and projects.

Knowledge is considered the key asset of modern organisations and, as such, industry and academia have been working to provide the appropriate support to leverage on this asset (Firestone & McElroy, 2003). Few examples of this work are: the extensive work on knowledge models and knowledge management tools, the rise of the so-called knowledge engineering area, the myriad of projects around 'controlled vocabularies' (such as ontologies, taxonomies, dictionaries, and thesauri), and the academic offer of knowledge-centred courses (graduation, master, doctoral).

The quest for innovation to be used as a wild card for economic development, growing and competitiveness, affects not only organisations, but also many countries. This demand for innovative processes and ideas, and the consequent pursuit of effectively more knowledge, raise inevitably issues regarding the adoption and use of Knowledge Management (KM) models and tools within organisations.

Knowledge Management theme and more specifically, how knowledge can be represented, gained a new dimension with the advent of the computer age. Particularly, with the creation of the World Wide Web, new forms of knowledge representation were needed in order to transmit data from source to recipient in common data formats, and to aid humans to find the appropriate answers for their questions in an easily understandable manner.

Artificial Intelligence (AI) based research abstracted knowledge into a clear set of parameters and used fairly static/rigid rules, had fairly limited "context" (the domain of its applicability), and were poor in "human communication." Further, such systems lacked interoperability because most AI tools focused on solving a specific problem and faced challenges with cross-context information flows, imputation, and interpretation, i.e., how to transfer the actual situation into the parameters used by the AI tool, (Dascal, 1992).

With the evolution of the Semantic Web, knowledge representation techniques came into the spotlight, aiming at bringing human understanding of the meaning of data to the world of machines. Such techniques create representations of Knowledge Sources (KS), whether they are web pages or documents (Figueiras, et al., 2012).

Like many Information Retrieval (IR) tasks, knowledge representation and classification techniques depend on using content independent metadata (e.g. author, creation date) and/or content dependent metadata (e.g. words in the document). However, such approaches tend to be inherently limited by the information that is explicit in the documents, which introduces a problem. For instance, in the situation where words like 'architect' and 'design' do not co-occur frequently, statistical techniques will fail to make any correlation between them (Nagarajan, et al., 2007).

Furthermore, existing IR techniques are based upon indexing keywords extracted from documents and then creating a term vector. Unfortunately, keywords or index terms alone often do not adequately capture the document contents, resulting in poor retrieval and indexation performances. Keyword indexing is still widely used in commercial systems because it is by far the most viable way to process large amounts of text, despite the high computational power and cost required to update and maintain the indexes.

Such challenges motivate the following question: how to intuitively alter and add contents to a document's term vector using semantic background knowledge available in domain ontologies, and thereby provide classifiers with more information than is exemplified directly in the document?

In the last decades, the use of ontologies in information systems has become more and more popular in various research fields, such as web technologies, database integration, multi agent systems, and Natural Language Processing. This work focuses on how ontologies can be used to improve semantic interoperability between heterogeneous information systems. We understand interoperability as the ability of two or more systems or components to exchange information and to use the information that has been exchanged (IEEE, 1990).

An ontology models information and knowledge in the form of concept hierarchies (taxonomies), interrelationships between concepts, and axioms (Noy & Hafner, 1997); (Noy & McGuinness, 2002). Axioms, along with the hierarchal structure and relationships, define the semantics, the meaning of the concepts. Ontologies are thus the foundation of content-based information access and semantic interoperability over the web.

Fundamentally, ontologies are used to improve communication between people and/or computers (Uschold & Jasper, 1999). By describing the intended meaning of "things" in a formal and unambiguous way, ontologies enhance the ability of both humans and computers to interoperate seamlessly and consequently facilitate the development of semantic (and more intelligent) software applications.

The motivation guiding this work is that a system should be interoperable and capable of wrapping existing data to allow for a seamless exchange of data among stakeholders - a necessary first condition for effective collaboration. Here, we propose to use background knowledge available in domain ontologies in to support the process of representing KS from the building and construction domain, thus improving the classification of such knowledge sources. In the scope of this work, ontology is a way to represent knowledge within a specific domain (Gruber, 1993).

Our hypothesis is that semantic background knowledge from ontologies can be used to the enrichment of traditional statistical term vectors can be fulfil by the usage of semantic background knowledge available in domain ontologies. Therefore, one of the main contributions of this work is consequently not trying to develop new or improving any of the current classification algorithms but to affect the document term vectors in a way that we could and measure the effect of such semantic enrichment on existing classifiers.

We believe that information contained in ontologies can be incorporated into many representation schemes and algorithms. In this paper, we focus on a particular representation scheme based on Vector Space Models (Salton, et al., 1975), which represents documents as a vector of their most important terms (knowledge representations). Important terms are those which are considered to be the best discriminators for each document space. The main aim is to understand how useful external domain knowledge is to the process of knowledge representation; what the trade-offs may be and when it makes sense to bring in such background knowledge. In order to do this, we intuitively alter basic *tf-idf* (term frequency–inverse document frequency) (Salton & Buckley, 1988) weighted document term vectors (statistic term vector) with the help of a domain ontology to generate new semantic term vectors for all documents to be represented.

This work presents the representation of KS through the use of Semantic Vectors (SV) based on the combination of the Vector Space Model (VSM) approach and a domain-specific Ontology (Costa, et al., 2012). Therefore, KS, in this work, are represented by SV which contain concepts and their equivalent terms, weights (statistical, taxonomical, and ontological), relations and other elements that semantically enrich each SV.

The performance of the proposed approach is evaluated using an unsupervised document classification algorithm. Document clustering has become one of the main techniques for organizing large volumes of documents into a small number of meaningful clusters (Chen, et al., 2010). However, there still exist several challenges for document clustering, such as high dimensionality, scalability, accuracy, meaningful cluster labels, overlapping clusters, and extracting semantics from the texts.

Also, performance is directly related with the quantity and quality of information within the Knowledge Base (KB) it runs upon. Until, if ever, ontologies and metadata (and the Semantic Web itself) become a global commodity, the lack or incompleteness of available ontologies and KBs is a limitation likely to have to be lived with in the mid-term (Castells, et al., 2007).

We used an unsupervised classification algorithm (K-Means clustering (MacQueen, 1967)) to evaluate the results of our approach. One of the reasons we choose an unsupervised classification is that supervised classification is inherently limited by the information that can be inferred from the training data. The objective here is to use a centroid-based document classification algorithm to assess the effectiveness of the altered vectors, due to the fact no in-depth knowledge of the actual contents of the document corpus was provided.

This paper is structured as follows. Section 2 illustrates a motivating scenario and the related work. Section 3 illustrates the domain ontology used under this work. Section 4 describes the process of enrichment of KSs. Section 5 illustrates the empirical evidences of the work addressed so far. Finally, section 6 concludes the paper and points to the future work to be carried out.

2. CHALLENGES

In order to understand the type of domain addressed within this work and the associated knowledge sources space, we present some of the relevant challenges in the B&C (Building and Construction) sector and why this topic is so important to this particular domain.

B&C projects are information-intensive. The availability of integrated project data and the recording of such data throughout the construction process are essential not only for project monitoring, but also to build a repository of historical project information that can be used to improve performance of future projects. This would allow construction actors to better share and use corporate knowledge when searching for appropriate actions to solve on-site construction problems. The shared knowledge is also expected to help better predict the impacts of corrective actions in a project life cycle, and so improve project performance.

The motivating scenario described here, corresponds to a realistic vision of the industry concerning the innovative way technology could be used to improve future collaborations. It describes a tactical meeting and illustrates some difficulties that can be met by architects when collaborating with other disciplines.

2.1 Motivating Scenario

Projects are conducted through a series of meetings and every meeting is considered a Decisional Gate (DG), a convergence point where decisions are made, problems are raised, solutions are (likely) found, and tasks are assigned to project participants. Pre-existing knowledge serves as input to the DG, the project is judged against a set of criteria, and the outputs include a decision (go/kill/hold/recycle) and a path forward (schedule, tasks, to-do list, and deliverables for next DG). The decisional gate representation is depicted in FIG. 1.

Each DG is prepared (through the creation of agendas), and the events that occur during the meeting shall be recorded. Between two DGs there is a permanent monitoring on the execution of all tasks executed. After meeting closure, there is a need for a mechanism to enable the preparation the minutes easily, highlighting the major decisions that were made during the meeting.

DGs normally go through the following phases: (i) Individual work; (ii) Initialisation; (iii) Collaboration; and (iv) Closing/Clean-up. Individual work relates to asynchronous collaboration, where all individuals involved in the project are supposed to provide inputs to the undergoing tasks. Initialisation (pre-meeting) covers the preparation of the meeting agenda and the selection of the meeting participants. Collaboration phase is the meeting itself where participants try to reach a common understanding regarding the issues from the agenda, using the right resources. This phase also considers the annotation of the decisions made during the meeting. Finally, Closing/Clean-up basically targets the creation of meeting minutes.

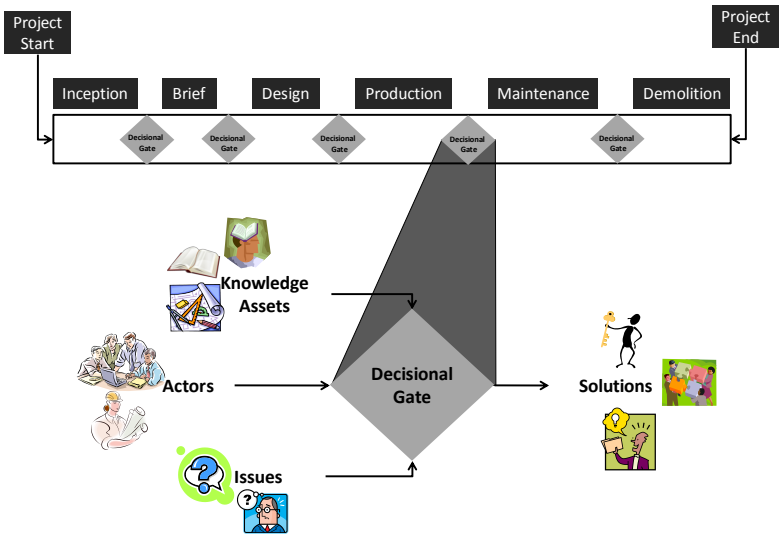

FIG. 1: The decisional gate

When considering the process, in the initial phase of a construction project, meetings need to be co-located as they permit the building of social links and trust between remote collaborators, i.e. across competencies or companies which is a prerequisite for efficient collaboration. Additionally, they enable the participants to share a common contextual understanding by enhancing in depth inter-disciplinary discussions. Indeed, they allow people to reformulate their discourse in order to adapt it to the level of understanding of other disciplines. The consequence is that better decisions can be made which consider and optimise several view points and that trust is increased between the collaborators/stakeholders. Finally, co-located meetings aim at formulating a consensus on a set of actions that have to be taken during or after the meeting as detailed previously which are then recorded (minutes). Subsequent meetings are held to resolve outstanding issues and address new ones. Actions held over potentially loose come of the contextual cohesion.

The example scenario described here, relates to a space that was originally designed to be a toilet for disabled people has been reduced in floor area. Indeed, the engineer had to include a separate installation shaft for supply and ventilation system in that space in order to respond to new requirements for fire protection and safety. As a consequence, the toilet has to be redesigned, but must include similar elements as previously planned: a close-coupled WC, a basin, a bathtub, a wall hung cupboard and a window (FIG. 2).

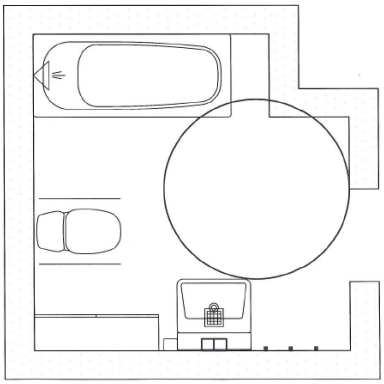

FIG. 2: A possible setting for the toilets' elements

The stakeholders are identified and invited to attend a meeting at the architectural company where the new proposed design must be presented and validated by a range of people with very different perspectives, interests and concerns. The identified stakeholders are presented in TABLE 1 and FIG. 3.

TABLE 1: Meeting stakeholders

Stakeholder	Function	Responsibility	Interests
Client	Building Owner or Investor	Issue Building Program and financing the project. Tender and contracting	Highest quality to lowest price - a.s.a.p.
Client Representative	Consulting	Securing the best possible result for the client. Building professional. All contractual obligations are fulfilled	Mediate client interest with the design and construction team
Project manager	Selected project leader for the project team	Depending on the type of tender used, he is overall contractual responsible from start to end. Legal approval and permits	Collaboration with the Client Representative. Communication of results from the design team.
Design manager	Selected project leader for the design team	Coordination of design team	Coordination between architects and engineers and possible other companies involved in the design phase
Architect	Architectural design	Control the look and feel of the building	Collaboration between the Design manager and the engineers
Engineer, mechanical	Design of heating, supply and sewage	Mechanical engineering	Collaboration between the Design manager and the architect
Engineer, structural	Design of building structure	Assure the structural integrity of the construction	Collaboration between the Design manager and the architect
Engineer, electrical	Design of electrical installations and communication	Assure the integration and integrity of the distribution network	Collaboration between the Design manager and the architect
Main Contractor	Building construction	Building quality, time and economy. Coordination of sub-contractors and suppliers.	To avoid conflicts with the other stakeholders and to complete with a profit
Supplier	Supply of materials	Deliver on time	To deliver with a profit
Municipal Architect	Represents the City Regulations and national building regulations	Secure that all national, regional, city and local area regulations and safety rules	Approve design and issue building permission
Council of Handicap	Represents people with special needs	Accessibility for all	Support to citizens with special needs
End-user representative	Tests and evaluate design – hands-on	Show functionality and introduce normal physical limits for persons with handicap	Create apartments for disabled persons with high quality despite personal physical limits.

The organisation of the meeting started a month before, in order to find common availability date of all the participants. It commenced with an email sent to all the main collaborators involved in the project. Each of them returned a list of companies that should be represented during the meeting. The second phase of the meeting organisation consisted in sending another email to all the relevant companies in order to identify the participants

that would be available for half a day on particular dates between the 23rd of April and the 4th of May. Each company returned a list of preferred dates and the details of the people who could attend the meeting on these particular dates.

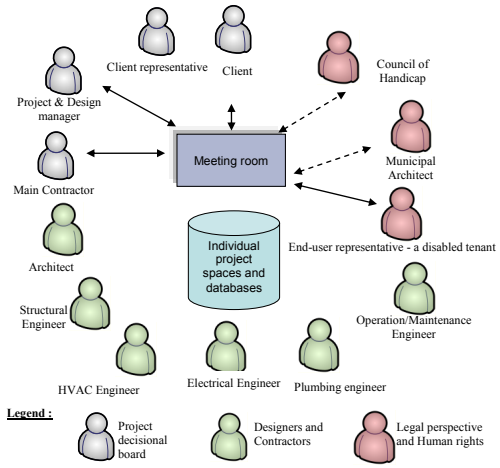

FIG. 3: The meeting participants, Current situation

The agenda prepared by the host company was sent by email to all the participants one week before the meeting. As a physical mock-up has to be tested during the meeting, the venue of the meeting has to be the storehouse of the contractor. The architect is the chair of the meeting. He starts with a Powerpoint presentation of the client's objectives and an explanation of the building design including a presentation of the reasons for constraints (FIG. 4 illustrates some measurements to be considered to design for disabled people) and a list of possible solutions for succeeding anyway. Then, an artistic 3D representation of the toilets is displayed and explained by the architect. Discussions start between the participants about some little issues due to the artistic representation of the toilets. Following this introduction, all the participants are invited to go to the physical mock-up to begin the discussions towards an agreement on the design modification between all the participants.

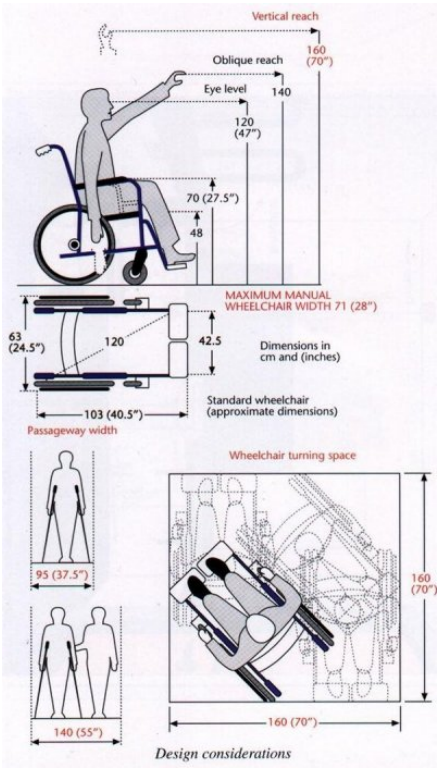

FIG. 4: Some measurement to be considered to design for disabled people (Couch & Forrester, 2003)

The mock-up has already been tested by the architect, the engineers and the main contractor in order to achieve a good level of certainty that the meeting would validate their proposition. However, artistic planning or structural engineering cannot design real human behaviour and preferences. Two disabled people are invited to test the usability of the toilets and that the elements are within reach and usable. While everyone is observing the design and trying to apprehend any future issue that could arise from it, the disabled person starts testing the mock-up. Immediately, she realises that the door is too narrow to enter with a wheel chair. Less than ten minutes later, it appears that there is not enough space in the room for using a wheel chair, that the window cannot be reached, and that some more space is required for any disabled people assistance.

Due to the fact that modifying a mock-up is a complicated task, the toilets have to be fully redesigned and a modified physical mock-up created before any validation is possible. The meeting ends with some discussions between all the participants in order to achieve a valid design for the next meeting. Discussions are limited because no back office verification is possible and there is only a flip-chart available for communication. The architect takes some notes on his notebook and quickly annotates some drawings. It is decided to exchange the bathtub with a shower arrangement - it cannot be decided what design should be used, but the end-users have some basic criteria such as even floor level, possibility to sit and panic button. At the end of the meeting, it is agreed that another two hour meeting will be scheduled approximately a week later with the same participants.

During the following meeting, a clash is found by the structural and mechanical engineers. Yet another meeting is needed, another week later, a total of 3 days spent on designing the toilet over a period of 3 weeks. The project cost is increased by the cost of the physical mock-ups and the travel expenses, as well as on the time and materials spent redesigning the toilet.

A possible futuristic scenario corresponds to the same situation as the one described above, but using of new technologies to improve co-located meetings. Its objective is to make the meetings more effective, which means that there is a better shared understanding between the participants, that more viewpoints can be considered and agreements and be resolved much faster. In order to achieve this, useful information has to be made available faster between all the participants independently of the location, in a way that is easily understood by people who need it. As a consequence, fewer meetings are required due to incomplete agreements, fewer problems have to be solved and the possibility to redesign as well as testing alternative solutions during the meeting. This fastens the building construction and makes the collaborators more available for fast responses in case their expertise is required for minor issues.

As in the previous project situation the project manager invites the relevant stakeholders to the collaborative workspace that are used for project meetings. A draft of an agenda is produced by the project manager and sent through the shared workspace to all the participants. Based on that agenda, the participants started selecting the tools they would need during the meeting and linked the documents considered relevant and data to a shared workspace that will be used during meeting. The approach described in the paper supports participants in this task, by helping them in identifying the most relevant documentation to be used during meeting based on the roles of the users and the constraints of the problem to be solved. The intention is to support meeting participants in identifying the most relevant sources, according to the problem constraints. Such early identification of knowledge sources will enable a better common understanding of the problem to be addressed, what can be used in order to promote more effective meetings, and also the recording of decisions taken after meeting closure. TABLE 2 points-out some relevant documentation topics, which needs to be identified by each participant prior to project meeting.

TABLE 2: Perspectives requirements for relevant documentation

Viewpoints	***Purpose***	***Sources of Knowledge***
Client	- control the cost - foresee problems - make the final decision	- 3D representation of the toilets - tender and contracts - extra costs/changes - up-to-date project budget - time to completion (Interest until sale/use) - sale and renting - legal documents - annotated toilets representations (measurements, components characteristics…) - contracts
Architect	- assure the uniformity of the design	- 3D representation of the toilets
Project Manager	- assure everyone's view is clearly expressed and understood - assure the constructability of the building - find the resources - control the cost - assure the feasibility of the tasks	- tender and contracts - extra costs/changes - up-to-date project budget - time to completion (Interest until sale/use) - sale and renting - 3D structural elements model - 3D plumbing elements model - 3D electrical elements model - 3D HVAC elements model - models annotations (dimensions, weight, assembly description) - tasks descriptions with workflow - people availability - roles and responsibilities (contractual requirements) - required tools - -supplied elements availabilities and prices
Engineer	(plumbing, maintenance…) - identify clashes - plan resources - identify future difficulties	- user characteristics (wheelchair database, anthropomorphic characteristics database…) - usage limitations (temperature, warranty…) - - interactive 3D representation of the toilets
Supplier	- share availability of the parts - share logistics - share prices - share products information	- tender and contracts - extra costs/changes - up-to-date project budget - time to completion (Interest until sale/use) - sale and renting - supplied elements availabilities and prices
Human rights	- assure the respect of the law - assure the respect of the user - control extravagance of the design	- 3D representation of the toilets - legal documents - annotated toilets representations (measurements, components characteristics…) - contracts
User	- test the design - identify usage difficulties	- 3D representation of the toilets - user characteristics (wheelchair database, anthropomorphic characteristics database…) - usage limitations (temperature, warranty…) - interactive 3D representation of the toilets - components characteristics - performance characteristics - maintenance characteristics
Participant	- express personal viewpoint and defend personal interests	- tacit knowledge

2.2 Related Work

As seen in the motivation scenario described above, knowledge needs to be shared in order to be proper capitalised during decision making processes. On one hand knowledge sharing is heavily dependent on technical capabilities and, on the other hand, since the social dimension is very strong during collaboration, there is also an increased need to take into account how to support the culture and practice of knowledge sharing. For instance, issues of trust are critical in collaborative engineering projects, since the distribution of knowledge and expertise means that it becomes increasingly difficult to understand the context in which the knowledge was created, to identify who knows something about the issue at hand, and so forth.

B&C knowledge is seen as a network of interoperable (evolving) models. Each model should be built by using a constructivist's epistemology. In other words, it should be built based on a bottom-up, field-centred, and human-oriented approach. These models or subdomain ontologies should be interlinked within a contemporary pragmatic approach. In other words, they should be integrated on the basis of utility to industry and usability and with the acceptance of the dual/relative nature of such models (ontologies) (El-Diraby, 2012).

A consensus strategy for interoperability embraces all standards where the main models of conceptualization are first created and subsequent data models are developed. Actors or developers harmonize their models with the intention of integrating their data models with other actors in the interoperability activity. This strategy consists of finding common concepts of the universe of discourse of the domain. In the case of the construction industry domain, the definition of those concepts is focused not only on construction products but also on construction processes during a project life cycle (ISO12006-3, 2006). The Industry Foundation Classes (IFC) captures specifications of actors, product, processes, and geometric representation, and provides support as a neutral model for the attachment of properties, classifications, and external library access (BuildingSmart, 2012). An example of separate international organizations that combine their efforts into a single object library is the International Framework for Dictionaries (IFD).

Human knowledge can be efficiently represented and shared through semantic systems using ontologies to encapsulate and manage the representation of relevant knowledge (Lima, et al., 2005). Specifically, ontologies provide knowledge conceptualization using a hierarchical system of concepts (taxonomies), associative relations (linking different concepts across hierarchies), and axioms (El-Diraby, et al., 2005). Thus, ontologies may enable reasoning about semantics between domain concepts and can play a crucial role in representing knowledge in the B&C industry (Lima, et al., 2005), (Rezgui, 2006).

A variety of semantic resources ranging from domain dictionaries to specialized taxonomies have been developed in the building and construction industry. Among them are BS6100 (Glossary of Building and Civil Engineering terms produced by the British Standards Institution); bcXML (an XML vocabulary developed by the eConstruct IST project for the construction industry); IFD (International Framework for Dictionaries); OCCS (OmniClass Classification System for Construction Information); BARBi (Norwegian Building and Construction Reference Data Library); and e-COGNOS (COnsistent knowledge management across projects and between enterprises in the construction domain). Within these semantic resources, the e-COGNOS project was the first to deploy a domain Ontology for knowledge management in the construction industry which has been tested in leading European construction organizations (Lima, et al., 2005).

The initiatives described are seen as efforts in order to establish a common ground for enabling semantic interoperability within the B&C sector. However many other web-based tools have used semantic systems to support some aspects of integrating unstructured data and/or ontologies. For example, the GIDS (Global Interlinked Data Store) technique distributes Linked Data across the network and then manages the network as a database (Braines, et al., 2009). The SWEDER mechanism (Semantic Wrapping of Existing Data Sources with Embedded Rules) makes available existing electronic data sources in a usable and semantically-rich format along with rules to facilitate integration between datasets (Braines, et al., 2008). The POAF technique (Portable Ontology Aligned Fragments) aims to support alignment between existing ontology resources. These techniques (and many others) can be used to create interoperability between interlinked unstructured data sets based on semantic analysis (Kalfoglou, et al., 2008). The Funsiec initiative employed IFC and taxonomies as conceptual models and starting points to create single, harmonized products, queries, and control vocabulary (Lima, et al., 2006).

For the sake of clarity, it makes worth to distinguish here the major difference between data and information. In this work, data is seen as a representation of the simplest facts about a system with limited meaningfulness. In information systems, data is normally stored in databases. Information is the composition of various data to establish a meaningful representation of facts. In information systems, information is exchanged normally through communication between humans or via electronic means such as web sites or e-mail. Typically, IT-based tools (such as XML and other web systems) are used to support the interoperable exchange of information.

The work presented here is a continuation of that in (Figueiras, et al., 2012) and (Costa, et al., 2012). Regarding the issue addressed in our work, Castells et al. (2007) propose an ontology-based scheme for the semi-automatic annotation of documents, and a retrieval system. The retrieval model is based on an adaptation of the classic vector-space model, including an annotation weighting algorithm. Similar to our approach Castells uses the *tf-idf* (term frequency–inverse document frequency) algorithm, matches documents' keywords with Ontology concepts, creates semantic vectors, and uses the cosine similarity to compare created vectors. However, Castells' does not take into consideration the nature and strength of relationships between concepts (either taxonomic or ontological) in a way that could influence performance on annotations, as we do.

The work presented by Sheng (2009) tries to overcome this drawback by presenting a way of mathematically quantifying hierarchical or taxonomic relations between ontological concepts, based on the importance of relationships and the co-occurrence of hierarchically related concepts, which can be reflected in the quantification of document semantic vectors. Sheng's work contributes by comparing the effectiveness of the traditional vector space model with the semantic model. Sheng used semantic and ontology technology to solve several problems that the traditional model could not overcome, such as the shortcomings of computing weights based on statistical method, the expression of semantic relations between different keywords, the description of document semantic vectors and quantifying similarity. However, Sheng's work neglects other types of semantic relations including ontological. According to Sheng's work, concept similarity decreases with the distance between concepts in a taxonomy, which seems not always the case as demonstrated with our approach. Sheng used 100 abstracts from document sources to evaluate his method; it would be interesting to use the full document texts in order to quantify how his approach scales up, when compared to the full document texts used by our approach. It should be mentioned that the approach used by Sheng's has been adapted and used in our approach to calculate the taxonomical relationship weights.

Another relevant approach in the area of IR and document classification is proposed by (Nagarajan, et al., 2007). The authors explore the use of external semantic metadata available in ontologies in addition to the metadata central to documents, for the task of supervised document classification. One of the key differences between Nagarajan's approach and ours is that Nagarajan does not quantify the difference between ontological related concepts and taxonomically related concepts. Also, our work does not directly include terms from documents within semantic vectors; the terms are first mapped to ontology concepts which guarantees a reduction in the semantic vector dimensionality and avoids a very sparse vector. A further key difference is that Nagarajan uses a supervised document classification algorithm, which is inherently limited by the information inferred from the training data as opposed to our approach of using an unsupervised clustering algorithm.

In other recent work, Xia et al. (2011) propose document classification mechanisms based on title vectors which assumes that the terms in titles represent main topics in those documents, and therefore the weights for title terms should be amplified. Xia's et al. (2011) work seems to show an improvement in text classification for webpages, where titles are carefully created by editors and usually reflect the main content of the webpage. However, the same does not apply to the technical documents considered in our work. As will be explained and demonstrated in later sections, document titles can sometimes be misleading about the real content of the document.

3. MODELLING THE BUILDING & CONSTRUCTION KNOWLEDGE

Models of B&C knowledge span three broad categories: classification systems and thesauri, product and process models, and ontologies. The first category is the most prominent and oldest. By far, these classification systems (such as the Swedish classification of construction terms - sfb, Uniclass and Masterformat) focused on product categorization with limited attention to ontological modelling. Product models such as IFC (Industry Foundation Classes) also have limited ontological features as they were geared towards assuring interoperable exchange of product data (in contrast to semantic knowledge).

International Framework for Dictionaries (IFD) is closely related to IFC and BIM (Building Information Modelling) and can be seen as a thesaurus of B&C terms with aims to create multilingual dictionaries or ontologies. It is meant as a reference library intended to support improved interoperability in the building and construction industry (BuildingSmart, 2012). The value of interoperability for a BIM-based construction projects has been analysed in Grilo and Jardim-Goncalves (Grilo & Jardim-Goncalves, 2010) and the authors support the conviction that interoperability in BIM can contribute to efficiency value levels, through supporting communication and coordination interactions between participants within BIM-based projects. The ontology developed under the scope of this work was intended to be IFC compliant and to capitalize on previous taxonomies/classification systems. BS6100 and UniClass terms were used to enrich the ontology.

From a high level point of view, the basic ontological model of the domain ontology was inspired by the e-COGNOS ontology (Lima, et al., 2005) and it can be described a as follows: a group of Actors uses a set of Resources to produce a set of Products following certain Processes within a work environment (Related Domains) and according to certain conditions (Technical Topics). As such, the proposed taxonomy includes seven major domains to classify these major concepts: Project, Actor, Resource, Product, Process, Technical Topics (Conditions), and Related Domains (work environment).

It is worth noting the first five domains coincide with major themes in the IFC model (FIG. 5) and the two other domains include related issues that are not covered by IFC. This ontology is process-centered. Other domains define all relevant process attributes. For example, the Technical Topics domain defines the concepts of productivity, quality standard and duration. The following subsections describe the major elements of these seven domains.

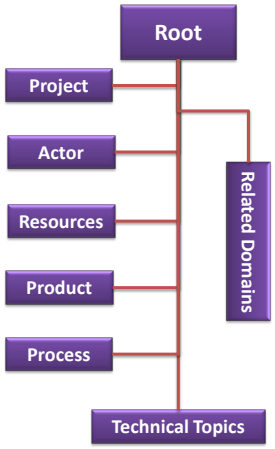

FIG. 5: Major domains in the domain ontology

All entities (including Process) have three ontological dimensions: state, stage and situation. State concept captures the status of entity development: dormant, executing, stopped, re-executing, completed. Stage concept defines various development stages: conceptualization, planning, implementation and utilization. Situation concept refers to planned entities and unplanned entities.

A Project is a collection of processes. It has two types: Brown field projects and Green field projects. It has a project delivery system, a contract, a schedule, a budget, and resource requirements. It also has a set of related aspects that include: start time, a finish time, duration, a quality standard, productivity level, a life cycle and a life cycle cost—all of which are defined in the Technical Topics domain.

A Process has an input requirements that include: the completion of all proceeding processes, the availability of required approvals, the availability of required knowledge items (documents, software, etc.), the availability of required Resources (materials, equipment, subcontractors), the availability of required Actors, and the availability of required budget. A Process has three major sub concepts: Phase, Activity and Task. It also has two major types: engineering process and administrative process. A Process has an output that include: update to a product time-line, an update to the project schedule, and update to the project budget, satisfaction/update to the legal conditions/status of Actors, may result in creating some project incidents (an accident, damage to an equipment, etc.).

A Product (also Actors, Processes and Resources) has attributes, parameters and elements, which are defined in Technical Topics.

The domain-specific Ontology used in this work was entirely developed using Protégé Ontology editor (Stanford Center for Biomedical Informatics Research, 2013), and is written in OWL-DL language (W3C, 2012). The Ontology comprehends two major pillars, namely concepts and their relations. The former relates to specific aspects (classes) of building and construction such as the type of project, project phase, geographical location and similar data. The latter specifies how the ontology concepts are related to each other.

Several levels of specificity are given for all families of concepts, as described for the 'Actor' concept. These specificity levels represent concept hierarchies and, ultimately, taxonomic relations such as 'Architect' <is_a> 'Design Actor' and 'Design Actor' <is_a> 'Actor'. All classes, or concepts, have an instance (individual), which corresponds to the class, and comprises the keywords or expressions gathered and related to each concept, through an ontological data-type property designated 'has Keyword'.

Concepts are related with a set of terms named 'equivalent terms' which are terms or expressions relevant for capturing different semantic aspects of such concepts. For instance, the 'Learning_Facility' concept has a 'Higher_Education_Facility' individual, and this individual has several equivalent terms such as 'university', 'science college' and 'professional college'. Thus each equivalent term belongs to some higher concept, as shown in FIG. 6. Moreover, concepts are connected by ontological object properties called 'ontological relations'. Ontological relations relate concepts among themselves and are described by a label (property) and the relevance (weight) of such relation in the context of the B&C domain Ontology.

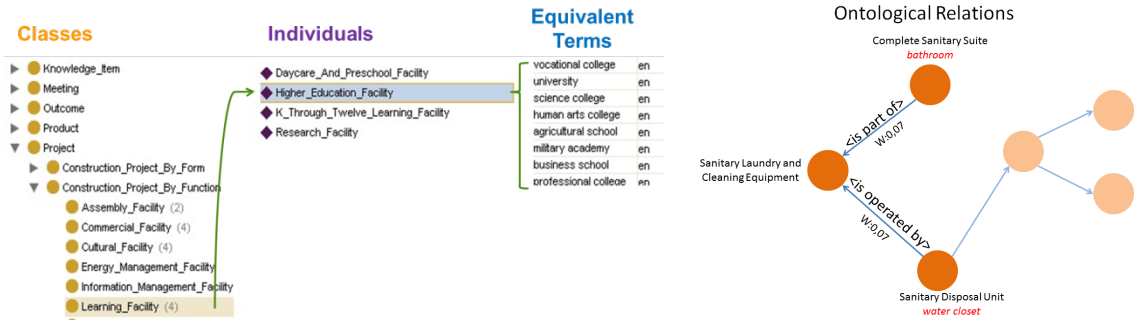

FIG. 6: Domain Ontology elements and relations

4. ENRICHING KNOWLEDGE REPRESENTATIONS PROCESS

In this section, we describe the rationale behind our hypothesis that semantic background knowledge from ontologies can be used to augment traditional statistical term vectors. Our approach mainly focuses on knowledge representation of knowledge sources, but there are several steps that need to be performed before and after the knowledge representation itself, as depicted in FIG. 7. The overall approach is described as follows:

- The first step deals with the searching of relevant knowledge sources, using the ICONDA digital library;
- The second step collects all relevant knowledge sources found, and stores them in a knowledge base repository;
- In the third step, knowledge experts within the B&C domain pre-label all relevant knowledge sources by inspection. This step will further be detailed under section 5;
- The fourth step is the core of our approach, which is detailed below in this section;
- The fifth step is responsible for applying an unsupervised classification algorithm (K-Means clustering), which groups knowledge sources into various categories (clusters). This step is further detailed in section 5;
- The final step evaluates the overall approach, using classical precision and recall metrics to measure performance. This step is also detailed in section 5.

FIG. 7: Step-wise approach

The core of our approach lies in altering document term vectors in three simple steps. FIG. 8 gives a general overview of the semantic vector creation process, which is carried out by three main modules, namely *Document Analysis Module*, and *Semantic Enrichment Module* (explained in sub-sections '4.1 Document Analysis Module', and '4.2 Semantic Enrichment Module', respectively). In our approach when receiving a set of textual documents, the document analysis module will extract terms, create the key term set, and produce a term occurrence statistical vector. From that point on, the semantic enrichment module will alter the statistical vector using information available in the B&C domain Ontology and produce an Ontology-concept occurrence vector or Semantic Vector for short.

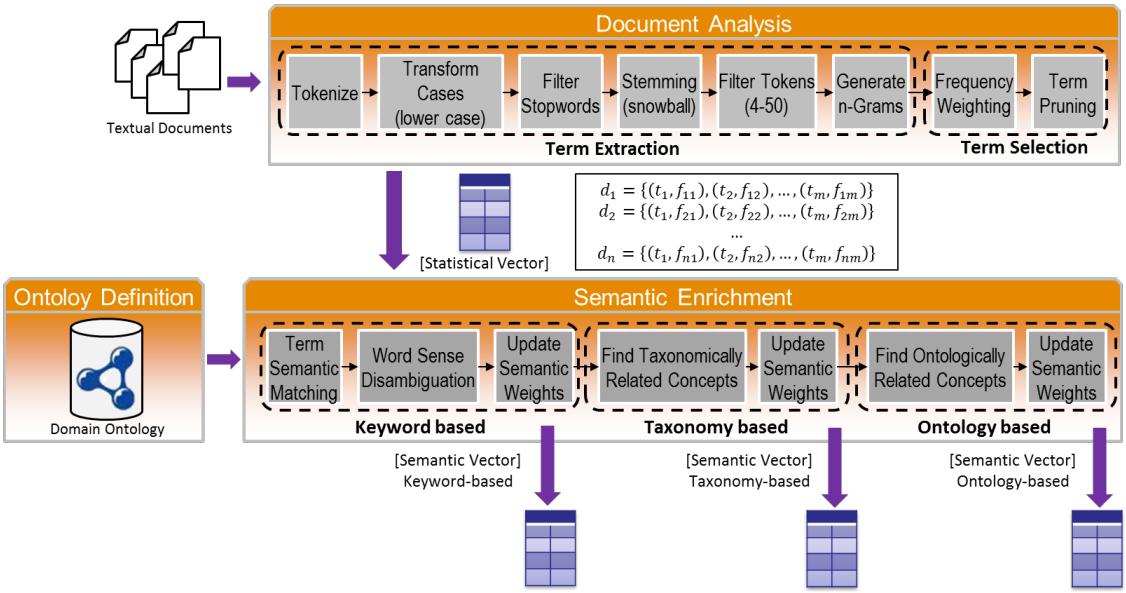

FIG. 8: The Semantic Vector creation process

4.1 Document Analysis Module

We start with a state-of-the art indexing tool, RapidMiner (RapidMiner, 2012), to generate document term vectors (statistical vectors) that order terms in a document by importance of their occurrence in that document and the entire document corpus by a normalized *tf-idf* score. There are two stages in this module, namely *Term Extraction* and *Term Selection*, which reduce the dimensionality of the source document corpus.

4.1.1 Term Extraction

The extraction process is as follows:

1. First, each document is split into sentences. Then, terms in each sentence are extracted as tokens (so called tokenization).
2. All tokens found in the document are transformed to lower case font.
3. Terms belonging to a predefined stop word list 1 are removed.
4. The remaining terms are converted to their base forms by a process called stemming, using the snowball method. Terms with the same stem are then combined for frequency counting. In this paper, a term is regarded as the stem of a single word.
5. Tokens whose length is "< 4" or "> 50" characters are discarded.

[1] It contains a list of stop words that is used by Rapidminer tool

6. The n-Grams generation is the creation of strings of 1 to N words. For this case we are considering the generation of unigrams (e.g. Energy), bigrams (e.g. Waste Management) and trigrams (e.g. Electric Power Product).

4.1.2 Term Selection

We consider that terms with low frequencies are most likely to be noise sources and of no use, so we apply the *tf–idf* (term frequency - inverse document frequency) method to select the key terms for the document set. Equation 1, is used for the measurement of $tfidf_{ij}$ for the importance of a term t_j within a document d_i. The main drawback of the *tf-idf* method is that long documents tend to have higher weights than short ones. The method considers only the weighted frequency of the terms in a document, but ignores the length of the document. To prevent this, in Equation 2, tf_{ij} is the frequency of t_i in d_j, and the total number of occurrences in d_j is the maximum frequency of all terms in d_j that is used for normalization to prevent bias for long documents.

$$tfidf_{ij} = tf_{ij} * idf_i \quad (1)$$

$$tf_{ij} = \frac{number\ of\ occurrences\ of\ t_i\ in\ d_j}{total\ number\ of\ occurences\ in\ d_j} \quad (2)$$

$$idf_i = \log \frac{number\ of\ documents\ in\ D}{number\ of\ documents\ in\ D\ that\ contain\ t_i} \quad (3)$$

After calculating the weight of each term in each document, those which satisfy a pre-specified minimum *tf–idf* threshold γ are retained. For this work, we consider only terms where the *tf-idf* score is ≥ *0.001* in order to reduce the high dimensionality of the generated vectors and also the computational power required to process the generated vectors. After close human inspection, it was concluded that terms which *tf-idf* score was less than 0.001, where not considered to be relevant enough. Subsequently, the retained terms form a set of key terms for the document set *D*.

A document, d_i, is a logical unit of text, characterised by a set of key terms t_j together with their corresponding frequency f_{ij}, and can be described in vector form by $d_i = \{(t_1, f_{i1}), (t_2, f_{i2}), \dots, (t_j, f_{ij}), \dots, (t_m, f_{im})\}$, the statistical vector. Thus for each document in the document corpus *D* there is a resultant statistical vector. A tabular example statistical vector is depicted in TABLE 3.

TABLE 3: Statistical Vector

Key Term	Weight
sanitari	0,004101
water_suppli_drainag	0,003265
Toilet	0,002482
personnel	0,002332

4.2 Semantic Enrichment Module

In this module we construct a new term vector, the Semantic Vector (SV) for all the documents in corpus *D*. This vector comprises Ontology concepts that are in the domain Ontology and whose equivalent terms (TABLE 4) semantically match terms which are present in the statistical vector (TABLE 2). This step ensures a 'meaningful' reduction in the term vector dimensionality and establishes a semantic grounding of the terms in the document that overlap with instances in the Ontology. However, there is a risk of obtaining a rather sparse vector if the domain ontology is itself sparse and poorly modelled. For now we assume the existence of a (relatively) complete ontology model.

Semantic vector creation is the basis for the approach in our work. It represents the extraction of knowledge and meaning from KS and the agglomeration of this information in a matrix form, better suited to mathematical handling than the raw text form of documents.

A semantic vector is represented by two columns: the first column contains the concepts that populate the knowledge representation of the KS, i.e., the most relevant concepts for contextualizing the information within

the KS; the second column keeps the degree of relevance, or weight, that each term has on the knowledge description of the KS.

TABLE 4: Ontological Equivalent Terms

Ontological Concept	Equivalent Terms
Complete_Sanitary_Suite	complete sanitary suite, complete bathroom suite, bathroom, washroom,…
Plumbing_Fixture_And_Sanitary_Washing_Unit	Bathtub, shower, service sink, lavatory,…
Sanitary_Disposal_Unit	water closet, toilet, urinal,…

Our approach takes into account three complementary procedures for creating the SV, where each procedure successively adds new semantic enrichment to the KS representation. The first step creates a keyword-based SV, the second step creates a taxonomy-based vector, and the final step creates an Ontology-based vector. Each step is described in the following sections.

4.2.1 Keyword-based semantic vector

The keyword-based SV takes into consideration only the liaison between terms present in the statistical vector and the concepts in the domain ontology. This step matches the statistical vector keywords with equivalent terms that are linked to the ontological concepts in the domain Ontology as shown in FIG. 9.

This process starts by first identifying the statistical vector keywords associated to a particular document and then finding similarities between each keyword and the equivalent terms within the ontology. The calculation of the similarities is done using the cosine similarity. The reason we choose the cosine algorithm is that cosine measure can be applied when comparing n-grams similarities of different magnitudes.

Cosine similarity algorithm measures the similarity between two vectors. In this case, we have to compare two n-grams. If we consider each one has a vector, we can use the cosine of the angle θ between x and y, represented in equation 4.

$$cos\ (x,y) = \frac{x}{\|x\|} . \frac{y}{\|y\|} \tag{4}$$

From equation1 in our study, this could be applied to our process in the following manner:

$$\frac{(Shared\ Keyword\ Terms)*(Shared\ Equivalent\ Terms)}{(Keyword\ Total\ Terms)*(Equivalent\ Terms\ Total\ Terms)} \tag{5}$$

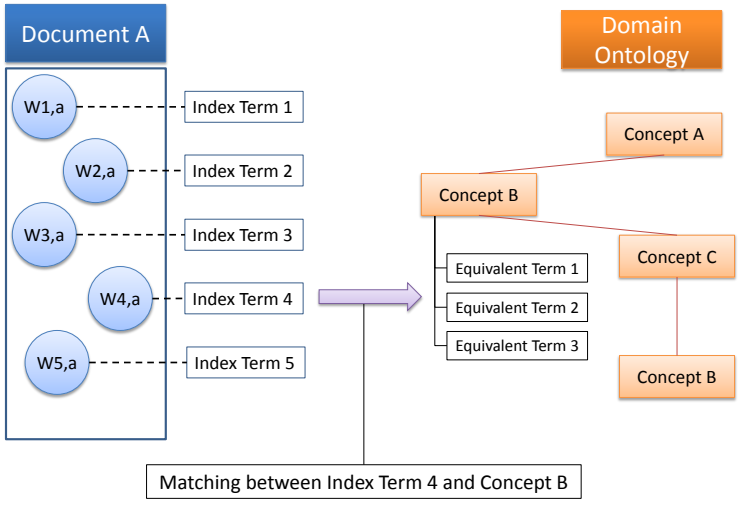

FIG. 9: Vector terms mapping against the Ontology concepts

Word sense disambiguation (WSD) for ontology learning, is a research topic which makes the matching process a challenge task. Most WSD research employs resources such as WordNet, text corpora, or social media. Many authors have proposed several approaches for dealing with the challenge of WSD (ex. (Wimmer & Zhou, 2013),

(Dandala, et al., 2013)). The implementation of a mechanism for word sense disambiguation is very relevant to the current scope of the work and the authors are considering it as part of future work. We came across several situations where word sense disambiguation is important, and at the moment is currently addressed through human inspection. FIG. 10 illustrates some examples of ambiguity found when creating an SV, where an equivalent term was inappropriately matched to a term in the statistical vector.

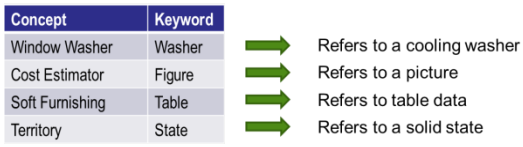

FIG. 10: Word Ambiguity Mismatch

Next the keyword-based SV is stored in the database in the form $\left[\sum_{i=1}^{n} x_i \; ; \; \sum_{i=1}^{n} w_{x_i}\right]$, where n is the number of concepts in the vector, x_i is the statistical representation of the concept and w_{x_i} is the semantic weight corresponding to the concept.

TABLE 5 depicts the weight of every ontological concept associated to each key term within the statistical vector, where the first column corresponds to the concepts that were matched to describe the most relevant terms extracted from the statistical vector shown in column 2, and the third column shows the semantic weight for each concept matched.

TABLE 5: Keyword-based semantic vector

Concept	Key Term	Weight
Sanitary_Disposal_Unit	toilet, urin, water_closet	0,149514
Sanitary_Laundry_and_Cleaning_Equipment_Product	sanitari	0,132629
Team	person, personnel	0,104497
Commitee	subcommitte	0,067880

4.2.2 Taxonomy-based semantic vector

Taxonomy-based vectors are the next step in the representation of KSs achieved by adjusting the weights of concepts according to the taxonomic relation among them, i.e., those concepts that are related by the 'is_a' type relation. If two or more concepts that are taxonomically related appear in a keyword-based vector, then the existing relation can boost the relevance of the expressions within the KS representation and therefore enhance weightings.

The taxonomy-based SV creation process defines a SV based on kin relations between concepts within the ontological tree. Specifically, the kin relations can be expressed through the notion of homologous/non-homologous concepts as follows (Sheng 2009).

Definition 1: In the hierarchical tree structure of the Ontology, concept A and concept B are homologous concepts if the node of concept A is an ancestor node of concept B. Hence, A is considered the nearest root concept of B, R(A,B). The taxonomical distance between A and B is given by:

$$d(A,B) = |depth(B) - depth(A)| = |depth(A) - depth(B)| \quad (6)$$

In Equation 4, depth (X) is the depth of node X in the hierarchical tree structure, with the ontological root concept depth being zero (0).

Definition 2: In the hierarchical tree structure of the Ontology, concept A and concept B are non-homologous concepts if concept A is neither the ancestor node nor the descendant node of concept B, even though both concepts are related by kin; If R is the nearest ancestor of both A and B, then R is considered the nearest ancestor concept for both A and B concepts, R(A,B). The taxonomical distance between A and B is expressed as:

$$d(A,B) = d(R,A) + d(R,B) \quad (7)$$

FIG. 11 depicts the difference between homologous and non-homologous concepts.

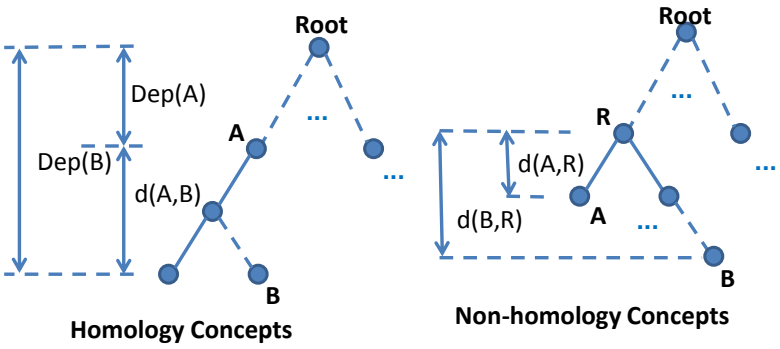

FIG. 11: Homologous and non-homologous concepts (Sheng, 2009)

The taxonomy-based SV is calculated using the keyword-based vector as input, where taxonomical relations are used to boost the relevance of the concepts already present within the vector or to add new concepts. The weight of the concepts is boosted when two concepts found in the keyword-based vector are highly relevant, with the degree of relevance being defined by a given threshold. If the relevance of the taxonomical relation between two concepts is higher than the predefined threshold, then the semantic weight of such concepts is boosted in the taxonomy-based vector. If a concept already present in the keyword-based vector is taxonomically related to a concept that is not present in the vector, then the related concept is added into the taxonomy-based vector.

One of the major differences between the present work and the work presented by (Sheng 2009) is that, in our approach, new concepts are only added into the taxonomy-based vector if the $d(A,B) = 1$ for homologous concepts and $d(A,B) = 2$ for non-homologous. The reason for such limitation is to avoid obtaining a sparse vector and to only add concepts that are highly related to already existing ones.

The intuition behind this work is to alter term vectors by strengthening the discriminative terms in a document in proportion to how related they are to other terms in the document (where relatedness includes all possible relationships modelled in an Ontology). A side effect of this process is the weeding out of the less important terms. Since ontologies model domain knowledge independently of any particular corpus, there is also the possibility of introducing terms in the term vector that are highly related to the document but are not explicitly present in it. The approach used for enhancing term vectors is therefore based on a combination of statistical information and semantic domain knowledge. An example of a taxonomy-based SV is depicted in TABLE 6.

The taxonomical similarity is calculated differently for both homologous and non-homologous taxonomical relations defined previously:

If $d(A,B) \neq 0$ and A and B are homologous.

$$Sim(A,B) = \left(1 - \frac{\alpha}{depth(A)+1}\right) \frac{\beta}{d(A,B)} \frac{son(B)}{son(A)} \quad (8)$$

If $d(A,B) \neq 0$ and A and B are non-homologous.

$$Sim(A,B) = \left(1 - \frac{\alpha}{depth(R)+1}\right) \frac{\beta}{d(A,B)} \frac{son(A)+son(B)}{son(R)} \quad (9)$$

If $d(A,B) = 0$

$$Sim(A,B) = 1 \quad (10)$$

TABLE 6: Taxonomy-based semantic vector

Concept	Weight
Sanitary_Disposal_Unit	0,107615
Sanitary_Laundry_and_Cleaning_Equipment_Product	0,092500
Team	0,075767
Plumbing_Fixture_and_Sanitary_Washing_Unit	**0,057912**

The concept 'Plumbing_Fixture_and_Sanitary _Washing_Unit' weight is boosted within the Taxonomy-based SV because it is highly related with the concepts 'Sanitary_Disposal_Unit' and 'Sanitary_Laundry_and_Cleaning_Equipment_Product'.

4.2.3 Ontology-based semantic vector

The third step in SV creation is the definition of the vector based on the ontological relations defined in the domain Ontology. We apply association rule theory to construct ontological concept relations and evaluate the importance of such relations for supporting the enrichment process of a domain ontology. The objective is to analyse the co-occurrences of concepts in unstructured sources of information in order to provide interesting relationships for enriching ontological structures. This is part of our on-going work described in (Paiva, et al., 2013).

The ranking of such semantic association is also complemented by human input (experts from the building and construction domain) to establish the final numerical weights on each ontological relationship. The idea behind having human intervention is to let the importance of relationships reflect a proper knowledge representation requirement, at first hand.

The creation of the ontological-based SV is a two-stage process using the taxonomy-based SV as input: the first stage boosts weights of concepts already present in the taxonomy-based vector, depending on the Ontology relations among them; the second stage adds new concepts that are not present in the input vector, according to ontological relations they might have with concepts belonging to the taxonomy-based vector (Costa, et al., 2012).

Analogous to the creation of a taxonomy-based SV, the new concept is added to the vector only if the importance of an ontological relation exceeds a pre-defined threshold, for the same constraint reasons. The ontological relation's importance, or relevance, is not automatically computed; rather, it is retrieved from an ontological relation vector comprising pairs of concepts and the weight associated to their relation, as shown in TABLE 7.

TABLE 7: Ontological Relations

Property	Subject	Object	Weight
is_part_of	Complete_Sanitary_Suite	Sanitary_Laundry_and_Cleaning_Equipment_Product	0,07
is_operated_by	Sanitary_Disposal_Unit	Sanitary_Laundry_and_Cleaning_Equipment_Product	0,07

Equation 9 describes the process of boosting of concepts or the addition of new ones, here Ow_{C_y}, is the new weight of the ontological concept, and Tw_{C_y} is the input taxonomy weight of the concept to be boosted. If the concept is added then Tw_{C_y} should be zero. Tw_{C_x} is the taxonomical weight of the concept related to C_y and $TI_{C_xC_y}$ is the weight of the relation between C_y and C_x.

$$Ow_{C_y} = Tw_{C_y} + \sum(all\ related\ C_xs)\left[Tw_{C_x} * \left(TI_{C_xC_y}\right)\right] \quad (11)$$

An example of an Ontology-based SV is depicted in TABLE 8.

TABLE 8: Ontology-based semantic Vector

Concept	Weight
Sanitary_Disposal_Unit	**0,111718**
Sanitary_Laundry_and_Cleaning_Equipment_Product	**0,099504**
Team	0,074115
Plumbing_Fixture_and_Sanitary_Washing_Unit	0,056649

In this example, the concepts 'Sanitary_Disposal_Unit' and 'Sanitary_Laundry_and_Cleaning_Equipment_Product' were boosted because they are already present in the taxonomy-based vector and are related by the ontological relation '<is_operated_by>'. On the other hand,

concepts ‘Team’ and ‘Plumbing_Fixture_and_Sanitary_Washing_Unit’, were not boosted, meaning that their respective weights were decreased after vector normalization.

5. ASSESSMENT OF THE PRESENTED WORK

This section describes the technical architecture of the prototype implemented to assess our approach and also the process for evaluating the results achieved so far.

5.1 Technical architecture

The architecture adopts a 3-tier model structure, comprising a knowledge repository layer, a service layer and a user interface layer. FIG. 12 illustrates the architecture, depicting also the technical modules addressed by each layer as well as the technologies used to develop the modules.

5.1.1 Knowledge repository Layer

The knowledge repository layer is composed by: (1) a document repository, developed under Liferay portal which is responsible for storing the all the Knowledge Sources that will be further processed; (2) a domain Ontology developed in OWL format and maintained developed by Protégé editor tool. A detailed description of the domain ontology was already provided in section 4.1; (3) and a relational database (named SEKS – Semantic Enrichment of Knowledge Sources) developed in MySQL, responsible for holding the appropriate statistical and semantic vectors for each of the knowledge source stored in the document repository. Meaning that, for each KS uploaded by the document repository portal corresponds to a set of vectors (statistical and semantic) stored in the SEKS database.

5.1.2 Service Layer

The service layer includes a set of web-services responsible for performing all the calculus needed for creating the semantic vectors associated to each KS, and also responsible for calculating the level of similarity between give a user query and such vectors. This layer is comprised by two different types of services, which are dependent on their level of nature:

The Basic Services consist of four service modules: Serialization Services, Calculus Services, Ontology Services, and Database Services. (1) Serialization Services are used by the Advanced Services and are responsible for converting messages that are exchanged between services to and from XML format. (2) Calculus Services are responsible for the required mathematical computations for creating the semantic vectors, and also the calculation of the similarity measure between two vectors using the cosine similarity algorithm. (3) Database Services are responsible for managing the ODBC connections and access from the service layer and the knowledge layer. (4) Ontology Services includes all necessary methods to access the elements of the domain Ontology, using the Jena API library, this enable to retrieve data form the OWL ontology.

The Knowledge Extraction Module, was developed using the RapidMiner tool, and enable to access the document repository and apply the *tf-idf* score to the document corpus, thus creating the statistical vectors for each document.

The Advanced Services layer interacts with all other basic services. It is responsible for performing the system’s main functionalities and comprises three high-level service modules: (1) Document Indexation Services handles all functions associated with the iterative creation of the three SVs as already explained in section 4.3. It takes as input the statistical vector created by the knowledge extraction module and as output creates the semantic vectors; (2) Query Treatment Services are responsible for transforming the user query into a semantic vector, and (3) Document Comparison Services contains all methods that support the comparison between the document corpus SVs and the user query. As output, this service presents a rank of the results of the comparison.

5.1.3 User Interface Layer

The user interface layer was developed using JSP, AJAX and JQuery technologies. It provides the front-end for the user to upload new documents into the document repository, navigate through the domain ontology and search for documents.

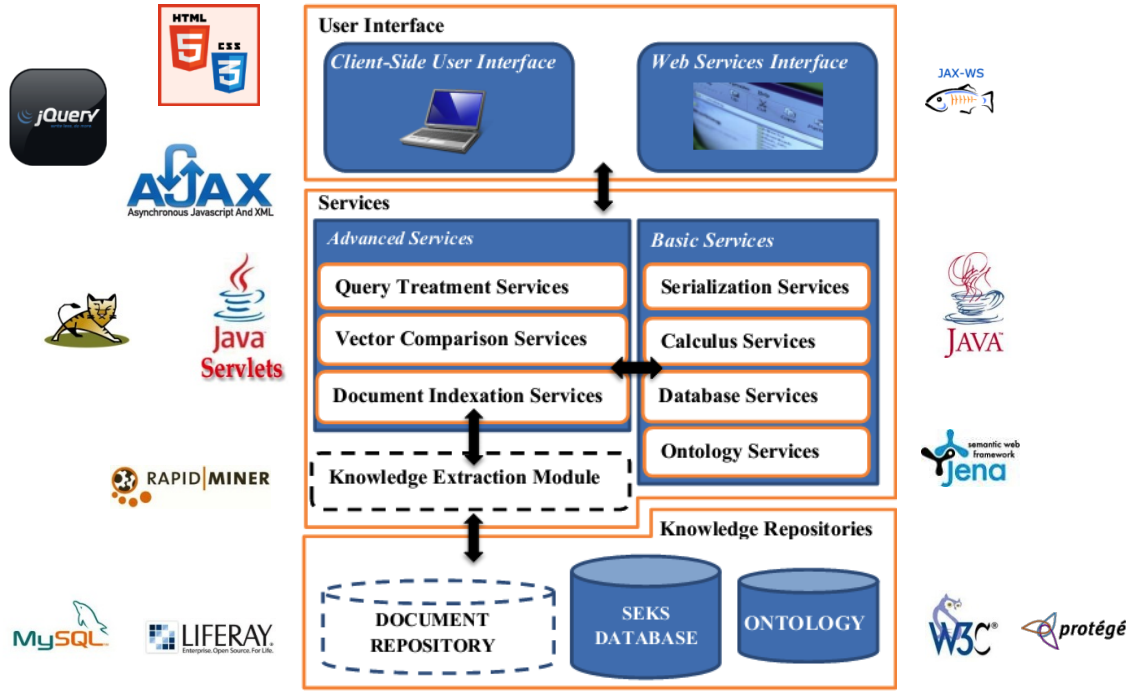

FIG. 12: Technical Architecture

5.2 Evaluation process

One of the key requirements for evaluating this approach is the availability of a relatively complete domain Ontology. This assessment built upon some preliminary results of prior work on semantic enrichment of knowledge sources (Figueiras, et al., 2012), (Costa, et al., 2012). A metadata knowledge base was developed focused on the building and construction sector, addressing the type of actors involved, related projects and products used.

Our dataset for evaluation in this paper is primarily focused on related products used in the building and construction domain. FIG. 13 shows part of the taxonomy into which the documents were classified. Although the taxonomy related to products we had available contained 16 sub-categories, we chose a smaller subset (5 categories as shown in FIG. 13) in order to analyse and explain the results in a clear fashion.

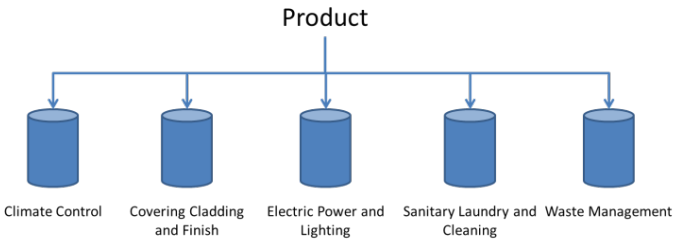

FIG. 13: Categories used for evaluation

We tested our approach with 20 scientific publications containing on average 3.500 words each. The reason for choosing scientific publications was the significant amount of words in each document, which makes the dispersion of each document regarding key terms much higher when compared to simple webpages or news headlines, and making the precise classification a greater challenge.

Documents used in the assessment were manually pre-labelled with the support of ICONDA search engine (IRB, 1986) and a close human evaluation, which sometimes helped in resolving some inconsistencies. For example looking into FIG. 14, ICONDA search engine considered such document into some extend related with 'lighting' concept, but after close inspection, such document was pre-labelled as 'climate control'.

- **TITLE:** ICT For Energy Efficiency: Towards Smart Buildings, Manufacturing, **Lighting** and Grids
- **Abstract:**...They are expected to have a significant impact on energy efficiency in the future. In this paper, the four industrial disciplines of buildings, manufacturing, **lighting** and power grids are identified to have great potential to deploy ICT to improve their energy efficiency...
- **Keywords:** ICT for energy efficiency, smart buildings, smart manufacturing, smart **lighting**, smart grids

FIG. 14: Pre-labelling mismatch example

The core aspect of our evaluation is to measure the effectiveness of the altered term vectors. The question we are trying to answer is whether our intuition of adding terms and boosting weights of terms in a term vector does, in practice, meaningfully amplify important terms and weed out less important ones? And at the same time, is it possible to represent knowledge sources with more accuracy with the support of domain ontologies? We believe that, having more accurate representations of knowledge sources can improve semantic interoperability among project teams, and consequently to facilitate knowledge sharing and reuse in B&C domain.

The comparison of this evaluation process is therefore performed between the four vectors – the statistical, keyword-based, taxonomy-based, and Ontology-based vectors.

As mentioned in earlier sections, the focus of this work is not on improving classification algorithms. Our system uses the altered term vectors as inputs to various classification algorithms - specifically, we used an unsupervised classification algorithm for the evaluations (K-Means clustering). The main reasons why K-Means clustering was adopted as an unsupervised classification algorithm is twofold: (i) its simplicity and low memory requirements; (ii) it gives best result when data set are distinct or well separated from each other.

5.2.1 The K-Means Clustering Algorithm

Let a set of text documents be represented as a set of vectors $X = \{X_1, X_2, \ldots, X_n\}$. Each vector X_j is characterized by a set of m terms $(t_1, t_2, \ldots, t_m)$. m is the total number of unique terms in all documents which form the vocabulary of these documents. The terms are referred to as features. Let X be a set of documents that contain several categories. Each category of documents is characterized by a subset of terms in the vocabulary that corresponds to a subset of features in the vector space.

A simple illustration of text data in VSM is given in TABLE 9. Here, x_j represents the jth document vector; t_i represents the ith term; each cell in the table is the frequency that term t_i occurs in x_j. A zero cell means that the term does not appear in the related document. Documents x_0, x_1, x_2 belong to one category C_0, assuming "Climate Control", while x_3, x_4, x_5 belong to another category C1, assuming "Waste Management".

Because these two categories are different, they are categorized by different subsets of terms. As shown in **Error! Reference source not found.**, category C_0 is categorized by terms t_0, t_1, t_2 and t_4 while category C_1 by terms t_2, t_3 and t_4. In the meantime, terms play different roles on identifying categories or clusters. For instance, the same frequency of t_4 appears in every document of category C_0, hence, t_4 should be more important than other terms in identifying category C_0.

TABLE 9: A Simple Illustration of Text Data in VSM

		t_0	t_1	t_2	t_3	t_4
C_0	x_0	1	2	3	0	2
	x_1	2	3	1	0	2
	x_2	3	1	2	0	2
C_1	x_3	0	0	1	3	2
	x_4	0	0	2	1	3
	x_5	0	0	3	2	1

K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let μ_k be the mean of cluster C_k. The squared error between μ_k and the points in cluster C_k is defined as

$$J(C_k) = \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (12)$$

$$J(C_k) = \arg\min_S \sum_{k=1}^{k} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (13)$$

Minimizing this objective function is known to be an NP-hard problem (even for K = 2). Thus K-means, which is a greedy algorithm, can only converge to a local minimum, even though recent study has shown with a large probability K-means could converge to the global optimum when clusters are well separated (Meilă, 2006). K-means starts with an initial partition with K clusters and assign patterns to clusters so as to reduce the squared error. Since the squared error always decreases with an increase in the number of clusters K (with J(*C*) = 0 when K = *n*), it can be minimized only for a fixed number of clusters.

- Supervised classification is inherently limited by the information that can be inferred from the training data (Nagarajan, et al., 2007). Meaning that, the accuracy and the representativeness of the training data, and also the distinctiveness of the classes must be taken into account. This tends to be a problem when dealing with large amounts document corpora, when no previous in-depth knowledge about the documents is assumed.
- Some documents tend to overlap, even when belonging to different categories. Such situations are quite common when working with documents with an average of 3.500 words each. In general, text classification is a multi-class problem (more than 2 categories). Training supervised text classifiers requires large amounts of labelled data whose annotation can be expensive. A common drawback of many supervised learning algorithms is that they assume binary classification tasks and thus require the use of sub-optimal (and often computationally expensive) approaches such as one vs. rest to solve multi-class problems, let alone structured domains such as strings and trees (Subramanya & Bilmes, 2008).
- Labelling such documents manually beforehand is not a trivial task and may affect adversely the training set of the classification algorithm. Our intention is to reduce as far as possible human intervention in the classification task and also to scale up our approach to hundreds of scientific publications.
- The goal of the assessment is to evaluate if the semantic enrichment process improves the similarity level among documents, even when such documents were not considered similar using purely statistical approaches but, indeed, they are in fact similar from a semantic perspective.

In the following sub-section, we present the results of our approach and give details of the kinds of classification patterns we have observed.

5.3 Results

Our metrics of evaluation are the traditional notions of precision and recall, and are computed as follows:

$$\text{Precision} = \frac{nº\ of\ documents\ correctly\ assigned\ to\ the\ category}{nº\ of\ documents\ correctly\ assigned\ to\ the\ category + nº\ of\ documents\ incorrectly\ assigned\ to\ the\ category} \quad (14)$$

$$\text{Recall} = \frac{nº\ of\ documents\ correctly\ assigned\ to\ the\ category}{nº\ of\ documents\ correctly\ assigned\ to\ the\ category + nº\ of\ documents\ incorrectly\ rejected\ from\ the\ category} \quad (15)$$

$$\text{Accuracy} = \frac{nº\ of\ documents\ correctly\ assigned\ to\ the\ category + nº\ of\ documents\ correctly\ rejected\ from\ category}{n} \quad (16),$$

where n = nº of documents correctly assigned to the category + nº of documents incorrectly assigned to the category + nº of documents incorrectly rejected from the category + nº of documents correctly rejected from the category.

Nevertheless, the correctness of the classification tends to be a subjective issue. What is a satisfactory classification for an application setting that has weighted ontological semantic relationships a certain way might

be unacceptable in other classification settings. The importance of relationships between ontological concepts is therefore an additional independent and tuneable component that affects the precision and recall metrics.

We first present some overall statistics and then discuss some success and failure patterns observed during correlation with the results of the classification. Tables 10 to 13, show average recall and precision values for 5 product categories comparing all four vectors.

TABLE 10: Performance using Statistical-based Vector

Accurracy:40%						
	True "Coating"	True "Waste Management"	True "Sanitary"	True "Lighting"	True "Climate Control"	**Class Precision**
Predicted "Coating"	2	0	0	0	0	**100%**
Predicted "Waste Management"	1	4	3	3	4	**26,67%**
Predicted "Sanitary"	0	0	1	0	0	**100%**
Predicted "Lighting"	0	0	0	1	0	**100%**
Predicted "Climate Control"	1	0	0	0	0	**0%**
Class Recall	**50%**	**100%**	**25%**	**25%**	**0%**	

TABLE 11: Performance using Keyword-based Vector

Accurracy:85%						
	True "Coating"	True "Waste Management"	True "Sanitary"	True "Lighting"	True "Climate Control"	**Class Precision**
Predicted "Coating"	4	0	0	0	0	**100%**
Predicted "Waste Management"	0	4	0	0	0	**100%**
Predicted "Sanitary"	0	0	2	0	0	**100%**
Predicted "Lighting"	0	0	0	3	0	**100%**
Predicted "Climate Control"	0	0	2	1	4	**57,14%**
Class Recall	**100%**	**100%**	**50%**	**75%**	**100%**	

TABLE 12: Performance using Taxonomy-based Vector

Accurracy:90%						
	True "Coating"	True "Waste Management"	True "Sanitary"	True "Lighting"	True "Climate Control"	**Class Precision**
Predicted "Coating"	4	0	0	0	0	**100%**
Predicted "Waste Management"	0	4	0	0	0	**100%**
Predicted "Sanitary"	0	0	2	0	0	**100%**
Predicted "Lighting"	0	0	0	4	0	**100%**
Predicted "Climate Control"	0	0	2	0	4	**66,67%**
Class Recall	**100%**	**100%**	**50%**	**100%**	**100%**	

TABLE 13: Performance using Ontology-based Vector

Accurracy:95%						
	True "Coating"	True "Waste Management"	True "Sanitary"	True "Lighting"	True "Climate Control"	**Class Precision**
Predicted "Coating"	4	0	0	0	0	**100%**
Predicted "Waste Management"	0	4	0	0	0	**100%**
Predicted "Sanitary"	0	0	3	0	0	**100%**
Predicted "Lighting"	0	0	0	4	0	**100%**
Predicted "Climate Control"	0	0	1	0	4	**80%**
Class Recall	**100%**	**100%**	**75%**	**100%**	**100%**	

As a result of looking more closely at some categories in order to understand the above results better, we discovered interesting patterns when the use of this approach added value and patterns when it did not.

Considering the 'Sanitary Laundry and Cleaning' category, we can conclude that using our approach there was a substantial improvement in terms of recall metric, from 25% using the statistical-based approach to 75% using the Ontology-based approach. In this case, the usage of ontological relations present in the domain Ontology (as shown in **Error! Reference source not found.**), improved the recall metric from 50% to 75%.

Our evaluation also indicated that quite a few documents had minimal or no direct matching with Ontology equivalent terms instances, mostly because of an incomplete domain ontology model (further investment in extending the Ontology knowledge base can address this issue to some extent) and the lack of a proper method for removing word ambiguity during the matching process (as explained previously).

It is possible for a domain Ontology to have no influence on the classification. Therefore the goal is to do no worse than the statistical-based approach whether the Ontology is relevant or wholly irrelevant.

Our dataset for evaluation considered (intentionally) several categories that had minor characteristic differences. For example, contents in 'Climate Control' and 'Electric Power and Lighting' categories have many similar predictor variables or terms that make classifying and allocating documents to the categories a challenge. Statistical term vectors that rely solely on document contents can rarely reliably classify a document as falling into one category or the other.

6. CONCLUSIONS AND FUTURE WORK

The paper's contribution targets the representation of KSs in various application areas for information retrieval, including, importantly, the semantic web. Moreover, it can also support collaborative project teams by helping them identify relevant knowledge amongst a panoply of KSs allowing knowledge to be better exploited within organizations and projects. Our contribution is at one hand highlight of the challenges of reconciling knowledge and to bring attention to the need for further research on the relationship of actors as social subjects and on the way how knowledge can be formalized and represented to the community. We anticipate the inclusion of this relationship in the research efforts will lead to a more effective sharing, exchanging, integrating, and communication of knowledge sources among actors through the employment of IT.

This work specifically addresses sharing and reuse of knowledge representation within collaborative engineering projects from the building and construction industry, adopting a conceptual approach supported by semantic services. The knowledge representations enrichment process is supported using a semantic vector holding a classification based on ontological concepts. Illustrative examples showing the process are part of this paper.

The intuition behind our work was to alter term vectors by strengthening the discriminative terms in a document in proportion to how strongly related they are to other terms in the document (where relatedness includes all possible relationships modelled in an Ontology). A side effect of the process was the weeding out of the less important terms. Since ontologies model domain knowledge independent of any document corpus, there is also the possibility of introducing relevant new terms into the term vector that are highly related to the document but not explicit in it.

The results achieved so far and presented here do not reflect a final conclusion of the proposed approach and are part of on-going work that will evolve and mature over time. Nevertheless preliminary results indicate that the inclusion of additional information available in domain ontologies in the process of representing knowledge sources can enrich and improve knowledge representations. Additional evaluation needs to be undertaken to reach more formal conclusions including additional metrics for assessing the performance of the proposed method. However, we can conclude that Ontologies do help improve the precision of a classification.

As described earlier, additional methods are required to reduce word ambiguity by taking account of context, when matching terms within the statistical vector with the equivalent terms present in the domain Ontology. At the moment the comparison is performed by using the cosine similarity algorithm, which may lead to inconsistencies as mention in earlier sections.

The domain Ontology is presently seen as something that is static and not evolving over time with organizational knowledge. The approach that is being exploitable is to extract new knowledge coming from KSs (new concepts and new semantic relations) and to reflect such new knowledge in the domain Ontology. The idea for accomplishing this is the adoption of association rules learning algorithms, correlating the co-occurrence of terms within the document corpus. Such measures can be considered as an estimation of the probability of terms being semantically related. The weights of such semantic relations should also be updated every time new KSs are introduced into the knowledge base corpus. The intent is that new ontological concepts and relations from new sources should be inserted and managed dynamically to support an evolving domain Ontology through a learning process.

7. REFERENCES

Braines, D. et al., 2009. *GIDS: Global Interlinked Data Store.* Hyattsville, International Technology Alliance.

Braines, D. et al., 2008. *A data-intensive lightweight semantic wrapper approach to aid information integration.* Patras, s.n.

BuildingSmart, 2012. *IFD Library for BuildingSmart.* [Online]
Available at: http://www.ifd-library.org/index.php?title=Home_Page
[Accessed 3 September 2012].

Castells, P., Fernandez, M. & Vallet, D., 2007. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. *IEEE Transactions on Knowledge and Data Engineering,* 19(2), pp. 261-272.

Chen, C.-L., Tseng, F. & Liang, T., 2010. An integration of WordNet and fuzzy association rule mining for multi-label document clustering. *Data & Knowledge Engineering,* pp. 1208-1226.

Costa, R. et al., 2012. *Capturing Knowledge Representations Using Semantic Relationships.* Barcelona, Spain, IARIA.

Couch, G. & Forrester, W., 2003. *Access in London: A Guide for People Who Have Difficulty Getting Around.* 4th ed. s.l.:Bloomsbury Publishing PLC.

Dandala, B., Mihalcea, R. & Bunescu, R., 2013. Word Sense Disambiguation Using Wikipedia. In: *Theory and Applications of Natural Language Processing.* s.l.:Springer Berlin Heidelberg, pp. 241-262.

Dascal, M., 1992. Why does language matter to artificial intelligence?. *Minds and Machines,* pp. 145-174.

Dumais, S., Platt, J., Heckerman, D. & Sahami, M., 1998. *Inductive learning algorithms and representations for text categorization.* Washington, ACM, pp. 148-155.

El-Diraby, T., 2012. Epistemology of Construction Informatics. *Journal of Construction Engineering and Management,* pp. 53-65.

El-Diraby, T., Lima, C. & Fiès, B., 2005. Domain Taxonomy for Construction Concepts: Toward a Formal Ontology for Construction Knowledge. *Journal of Computing in Civil Engineering,* 19(4), pp. 394-406.

Figueiras, P. et al., 2012. *Information Retrieval in Collaborative Engineering Projects – A Vector Space Model Approach.* Barcelona, Spain, INSTICC, pp. 233-238.

Firestone , J. & McElroy, M., 2003. *Key Issues in the New Knowledge Management.* Burlington: Butterworth-Heinemann.

Grilo, A. & Jardim-Goncalves, R., 2010. Value proposition on interoperability of BIM and collaborative working environments. *Automation in Construction,* p. 522–530.

Gruber, T., 1993. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies,* pp. 907-928.

IEEE, 1990. *Standard computer dictionary – a compilation of IEEE standard computer glossaries.* The Institute of Electrical and Electronics Engineers: s.n.

IRB, F., 1986. *ICONDA®Bibliographic.* s.l.:s.n.

ISO12006-3, 2006. *Building construction - organization of information about construction works - Part 3: Framework for object-oriented information,* Switzerland: International Organization for Standardization.

Kalfoglou, Y., Smart, P., Braines, D. & Shadbolt, N., 2008. *POAF: Portable Ontology Aligned Fragments.* Tenerife, s.n.

Lima, C., El-Diraby, T. & Stephens, J., 2005. Ontology-based optimisation of knowledge management in e-Construction. *ITcon,* Volume 10, pp. 305-327.

Lima, C., Silva, C., Duc, C. & Zarli, A., 2006. A Framework to Support Interoperability among Semantic Resources. In: *Interoperability of Enterprise Software and Applications.* s.l.:Springer London, pp. 87-98.

MacQueen, J., 1967. *Some methods for classification and analysis of multivariate observations.* Berkeley, University of California Press.

Meilă, M., 2006. *The uniqueness of a good optimum for K-means.* Pittsburgh, ACM, pp. 625-632.

Nagarajan, M. et al., 2007. *Altering Document Term Vectors for Classification - Ontologies as Expectations of Co-occurrence.* Alberta, ACM, pp. 1225-1226.

Nonaka, I. & Takeuchi, H., 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation.* New York: Oxford University Press.

Noy, N. F. & Hafner, C., 1997. The State of the Art in Ontology Design. *AI Magazine,* pp. 53-74.

Noy, N. & McGuinness, D., 2002. *Ontology Development 101: A Guide to Creating Your First Ontology,* Stanford: Knowledge Systems Laboratory.

Paiva, L., Costa, R., Figueiras, P. & Lima, C., 2013. *Discovering Semantic Relations from Unstructured Data for Ontology Enrichment - Association rules based approach.* Lisbon, IEEE.

RapidMiner, 2012. *Rapid-I GmBH.* s.l.:s.n.

Rezgui, Y., 2006. Ontology-Centered Knowledge Management Using Information Retrieval Techniques. *Journal of Computing in Civil Engineering,* 20(4), pp. 261-270.

Salton, G. & Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management,* pp. 513-523.

Salton, G., Wong, A. & Yang, C. S., 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM,* November, 18(11), pp. 613-620.

Sheng, L., 2009. A Semantic Vector Retrieval Model for Desktop Documents. *Journal of Software Engineering and Applications,* 2(1), pp. 55-59.

Stanford Center for Biomedical Informatics Research, 2013. *Stanford's Protégé Home Page.* [Online]
Available at: http://protege.stanford.edu/
[Accessed 3 Spetember 2012].

Subramanya , A. & Bilmes, J., 2008. *Soft-supervised learning for text classification.* Honolulu, Hawaii, Association for Computational Linguistics, pp. 1090-1099.

Uschold, M. & Jasper, R., 1999. *A Framework for Understanding and Classifying Ontology Applications.* Stockholm, CEUR Publications.

W3C, 2012. *OWL Web Ontology Language Reference.* [Online]
Available at: http://www.w3.org/TR/owl2-overview/
[Accessed 2012 September 3].

Wimmer, H. & Zhou, L., 2013. *Word Sense Disambiguation for Ontology Learning.* Chicago, s.n.

Xia, T. & Du, Y., 2011. *Improve VSM Text Classification by Title Vector Based Document Representation Method.* Singapore, IEEE.