A Methodology for Structured Ontology Construction applied to Intelligent Transportation Systems

D. Gregor[a], S. Toral[b], T. Ariza[c], F. Barrero[b], R. Gregor[d], J. Rodas[d], M. Arzamendia[a]

[a] *Laboratory of Distributed Systems, Faculty of Engineering, National University of Asuncion, 2060 Isla Bogado, Luque, Paraguay*
[b] *Department of Electronics Engineering, University of Seville, 41092 Seville, Spain*
[c] *Department of Telematic Engineering, University of Seville, 41092 Seville, Spain*
[d] *Laboratory of Power and Control Systems, Faculty of Engineering, National University of Asuncion, 2060 Isla Bogado, Luque, Paraguay*

Abstract

The number of computers installed in urban and transport networks has grown tremendously in recent years, also the local processing capabilities and digital networking currently available. However, the heterogeneity of existing equipment in the field of ITS (Intelligent Transportation Systems) and the large volume of information they handle, greatly hinder the interoperability of the equipment and the design of cooperative applications between devices currently installed in urban networks. While the dynamic discovery of information, composition and invocation of services through intelligent agents are a potential solution to these problems, all these technologies require intelligent management of information flows. In particular, it is necessary to wean these information flows of the technologies used, enabling universal interoperability between computers, regardless of the context in which they are located. The main objective of this paper is to propose a systematic methodology to create ontologies, using methods such as a semantic clustering algorithms for retrieval and representation of information. Using the proposed methodology, an ontology will be developed in the ITS domain. This ontology will serve as the basis of semantic information to a SS (Semantic Service) that allows the connection of new equipment to an urban

*Corresponding author
Email address: dgregor@ing.una.py (D. Gregor)

network. The SS uses the CORBA standard as distributed communication architecture.

Keywords: Intelligent Transportation Systems, Ontology, Clustering, Information Retrieval, Collaboration, CORBA, Distributed Systems, Statistical Data Analysis.

1. Introduction

The real-time estimation of traffic parameters and the control operations constitute a challenge for control of urban traffic systems (Chen and Cheng, 2010). Until now, the equipments installed in urban networks usually work in a centralized way, providing information to the traffic control center through the urban data network and performing actions according to the decisions of an operator at the control center. However, the enhancements of transport equipments due to the evolution of electronics and data networks allow them to share information and work cooperatively. The main challenge in the design and operation of ITS is information exchange, which is a difficult task in highly distributed systems. From a technical standpoint, there are difficulties in integrating information using compliant standards and connecting multiple systems, especially when considering the complexity and volume of information flows involved in the field of ITS, where both, the hardware as the data generated are highly heterogeneous (Toral et al., 2010). Therefore it is necessary to optimize the interoperability, security and efficiency of processes and devices which are part of the ITS by developing new technologies. More specifically, it is necessary to analyze the needs of the transport and logistics from a multimodal perspective, and to design new systems and tools able to provide "higher intelligence" in the process of information exchange and interoperability between devices. Distributed systems are well known for their difficulty of interoperation among agents, which justifies the interest in unified software platforms (Wang et al., 2006). SOA (Service-Oriented Architecture) is presented as an attractive alternative to enable interoperability of systems and the reuse of resources. But SOA applications face many security problems during design and development (Qu et al., 2010). In SOA architectures, the WS (Web Services) are a commonly used technology. WS use SOAP (Simple Object Access Protocol) as the communication protocol between various services. SOAP is an XML-based protocol. However, processing large SOAP messages significantly reduces system performance, causing bottlenecks in comparison with other technologies like CORBA (Tekli et al., 2012). This represents a problem in wireless communication networks (Phan et al., 2008) and in the ITS field, where the number of connected devices is growing over time. In practice, SOA-based applications are not always successful as most of them are done on an ad-hoc basis, and primarily based on personal experiences (Guo et al., 2010). Although companies are increasing their dependence towards SOA, these systems are still in an immature early stage with important security problems (Kabbani et al., 2010). The common problem in all the mentioned technologies is the interoperability between services and devices that are part of ITS, due to the differences in the information representation and semantics. The use of ontologies in this field would be a solution to enable reuse of domain knowledge and to generate smart clients. Agents that share semantic information could use this ontological information to respond to requests DD (Device-Device), serve as input to other services, enable reuse of domain knowledge or work cooperatively with other existing ontologies.

A methodology to define an ontology in the field of ITS is proposed. The ontology will be used in a CORBA-compliant Semantic Service, which allows finding services in a distributed environment. The developed ontology will serve as initial DataBase to the intelligent system of semantic management, where the hardware devices can exchange information through a communication system and work cooperatively. Section 2 provides an overview of previous related work. In Section 3, the proposed methodology is introduced, using as a starting point Systematic Literature Review (SLR) techniques, and then applying semantic analysis techniques and statistical data analysis to build the ITS ontology. Section 4 details the obtained results with the proposed methodology, testing the resulting ontology in a CORBA distributed environment. Finally, Section 5 shows the conclusions and future work.

2. Related Work

One of the main challenges in the ITS is the cooperative traffic. The idea of cooperation within ITS was initiated by the concept of cooperative in automated highways where vehicles receive input signals from the road environment. The first ideas documented on automated highways were presented in 1960 by the research laboratory of the General Motors (Gardels, 1960). A cooperative traffic system makes use of data as soon as they are collected, automating decision making in situations that require the intelligent intervention of ITS environment. (Soares et al., 2009) present a strategy in data dissemination for cooperative systems, defending that diffusion policies plays a determining role in the spread of ITS for the efficient information propagation. Indeed, the main objective of the cooperative driving is to focus on prevention and early detection of risks. However, this study does not specify how to find or maintain information. (Rockl and Robertson, 2010) argue that the success of cooperative ITS applications is mainly affected by the exchange of information between distributed nodes. According to authors, the transmission of large amount of information contrasts to the limited bandwidth of the channels that tend to be shared by all nodes participating in the ITS. But the extraction and interpretation of the information is out of the scope of the study. Therefore, it is necessary to develop efficient heterogeneous alternatives to increase the effective capacity of the ITS and to improve the efficiency of the transport systems. The solution lies mainly in the cooperative commitment to select relevant pieces of information for dissemination according to their value.

With the increasing development of electronics and the possibility of using embedded systems with increasing processing capabilities, the concept of cooperation has been extended from the original idea of cooperative driving to the current ITS distributed systems. The main idea of cooperation in ITS distributed systems is based on the collaboration of vehicle driving with available services in urban, suburban, metropolitan and rural areas, where vehicles interact with the environment, and the environment itself acts intelligently based on traffic events. (Mitropoulos et al., 2010) presented a system called WILLWARN (Wireless Local Danger Warning) based on recent and future trends in cooperative driving allowing electronic security to prevent risks through "Vehicle-Hazard" detection applications on-board, V2V (Vehicle to Vehicle) and V2I (Vehicle to Infrastructure) communications. One of the main causes of road accidents is the excessive and slow reaction of the driver in critical situations. However, the system proposed by Mitropoulos is exclusively focused on managing messages alerting the driver of the danger in ad-hoc basis, ignoring the quality and presentation of information.

(Thomas and van Berkum, 2009) proposed a prediction scheme for recurring traffic events based on data collected at urban intersections. They argue that it is necessary the management of events on demand in case of possible incidents, but they do not validate the results of the analysis with real data of incident detection, and they do not define how the information is collected, shown or stored. The main challenges in current ITS distributed architectures, where information plays an important role, are the heterogeneity of software, hardware devices, and communication networks. In the case of hardware devices it is usual the incompatibility in the data representation, the problems of synchronization and the wide variety of controllers and processors. Software applications and services have problems caused by the existence of multiple programming languages, different versions of the same application or service, the competition between proprietary and free/open source software as well as problems of understanding and distributed DataBases complexity. Finally, heterogeneity in communication networks is mainly due to the wide variety of network protocols, and the deployment of distributed networks, in some cases incompatible with traditional networks.

To overcome these drawbacks, ontologies can be an important issue in the future of ITS. One of the main advantages of the integration of ontologies in ITS is the intelligent and secure semantic location of services with certain characteristics and properties. From the point of view of interoperability between devices from different vendors and platforms, the most striking advantage is the intelligent information retrieval. Services can be published in descriptive ontologies and devices can make use of data and metadata from different kinds of runtime traffic events. While more structured is the services information, more accurate, fast and smart they can be found. Metadata can provide some semantics to this problem since ontologies provide a conceptual framework to exploit through metadata exchange schemes. Numerous previous studies have made use of metadata to improve implementation of collaborative applications in different scenarios. (García et al., 2012) present a context model based on an ontology which takes a combined approach to model the context information used by transport services.

The modeled distributed information is related to a primary context about the location, time, identity and quality of services, but applied only to a service for location of parking spaces. Thanks to the proposed scenario, they demonstrate that context information generated from autonomous distributed sources can be represented using a common data model and can be structured according to a common ontology. The resulting data can be shared, associated, fused, or reasoned. (Chen et al., 2008) proposed the design and implementation of a framework for public transport. They include a mechanism for data collection through WS which are specifically used for planning routes. However, the use of WS usually based on SOAP and XML may cause excessive bandwidth consumption for more complex systems where there is a big demand for services. (Fernandez and Ossowski, 2011) support the assumption that the use of MAS (Multi Agent Systems) enables a decoupled design and the implementation of different modules (agents), encouraging reuse of similar ontological domains, reducing the development effort and increasing system reliability (reuse of existing services). They focus the study on a service oriented multi-agent architecture for constructing advanced DSSs (Decision Support Systems) in transport management. However, they do not specify how to use the information as a tool or how to work cooperatively with other existing ontologies. (Terziyan et al., 2010) detail the requirements and the necessary architecture for traffic management systems, showing how such a system can be beneficial from the semantic point of view through technologic agents but questioning how this system can be combined with data processing and automated tools. A system for information retrieval based on a fuzzy ontological framework was proposed by (Zhai et al., 2008). The proposed framework is composed of three elements: concepts, properties of concepts and values of properties, being the property value any standard data type or linguistic values of fuzzy concepts. The main drawback of this framework is that the information retrieval system is primarily focused on information about traffic accidents, leaving aside other key issues such as interoperability between devices or heterogeneity of information. A cooperative traffic system should be able to solve complex problems using environmental data and metadata. The ITS equipments should be prepared to learn from the environment and change their characteristics based on events. Additionally, they must be able to interact with each other, forming multi-agent systems to achieve objectives. In this paper it is proposed an intelligent solution in the recovery and management of heterogeneous information in order to build an ontology using a taxonomy as the starting point of the study. The ontology will serve to organize and offer a metadata based service spread across the traffic network.

One of the first steps in the ontology construction is undoubtedly the IR (Information Recovery). Due to the large amount of available information, building ontologies from scratch and manually would require a lot of time and effort. Therefore, it is necessary to incorporate scientific techniques in the analysis and dynamic selection of information to provide a logical structure. Scientific Systematic Literature Review (SLR) is the field of study that tries to analyze and integrate essential information of the primary research studies on particular topic, in a perspective of set unitary synthesis. SLR has become an important research methodology for the recovery and collection of information (Hall et al., 2012). The aim of SLR is the identification, evaluation and interpretation of all relevant research studies about a particular research question using rigorous methods and specific algorithms. (Zhang et al., 2011) argue that the accuracy and preciseness in the information search process is actually a critical point that distinguishes systematic reviews from the traditional ad-hoc literature reviews. They have developed a systematic approach based on evidence for the development and implementation of optimal search strategies on digital literature. The proposed approach incorporates the concept of "quasi-gold standard" (QGS), which is the collection of known studies, and the corresponding "quasi-sensitivity" in the search process to evaluate its performance. There are several works about methodologies for developing ontologies. (Gruninger and Fox, 1995), proposed a methodology to design and evaluate an ontology that first intuitively identifies the possible applications where the ontology can be used. They use a set of questions called "competency questions" to determine the scope of the ontology and to extract key concepts, properties, relations and axioms. A more systematic approach for the construction of an ontology from scratch is the so called Methontology (Fernandez et al., 1997). This is perhaps one of the most complete proposed methods and considers the development of ontologies as a computer project. It includes activities for project planning, quality results, documentation, etc., and allows the building of new ontologies or reusing existing ones. (Chandrasegaran et al., 2013) applied a formal concept analysis methodology to develop a domain-specific ontology. They used a formal concept analysis to identify similarities among a finite set of objects based on their properties, providing a conceptual hierarchical clustering. However, the above methods lack of tools for IR and SLR. It is necessary to consider IR and SLR as part of the methodology for ontology creation in order to avoid bias in the resulting ontology, as the methodology proposed in this paper.

3. METHODOLOGY

Fig. 1 shows a block diagram of the proposed methodology for developing the ontology. The main objective of the proposed methodology is to discover ITS services based on common patterns among the data, with the final aim of obtaining class hierarchies in the "Building the Ontology" block.

The proposed methodology includes several automated methods for developing meta-analysis techniques on documents. The starting point is a taxonomy that summarizes the main topics of a domain field and a collec-

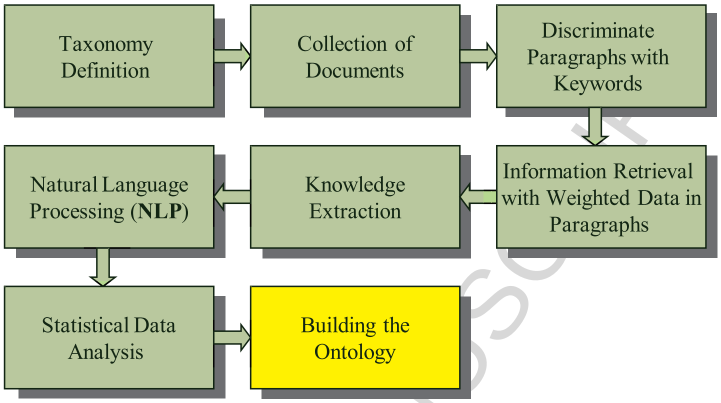

Taxonomy Definition
Collection of Documents
Discriminate Paragraphs with Keywords
Information Retrieval with Weighted Data in Paragraphs
Knowledge Extraction
Natural Language Processing (**NLP**)
Statistical Data Analysis
Building the Ontology

Figure 1: Block Diagram: Proposed Methodology.

tion of documents representing the major research trends in the ITS area. Based on the proposed methods, a complete ontology of services and service containers in the domain of ITS can be built. The results are a conceptual scheme that can be exploited through metadata exchange among devices and embedded applications in distributed urban systems. The following subsections describe in detail each block listed in the general scheme of the proposed methodology.

3.1. Taxonomy Definition

The first step for developing an ontology consists of obtaining a set of basic concepts or classes that define a specific domain, the ITS field in this case. Typically, this step involves the search of a set of keywords covering all the topics and issues related to the target domain. However, in the case of the ITS field, several organizations like U.S. DOT (United States Departments of Transportation) have previously explored this field in detail (RITA U.S. DoT, 2015). More specifically, the Research and Innovative Technology Administration (RITA) coordinates the U.S. Department of Transportation's research programs and it is in charge of the advances in the deployment of cross-cutting technologies to improve the transportation system (RITA, 2015), (USDOT, 2015). As part of their activities, they have developed a taxonomy of the ITS field considering several Levels of Detail (LoD), as shown in Fig. 2, which represents a part of the RITA U.S. DoT taxonomy. In this study, it has been considered this taxonomy until the LoD 4, which provides a collection of 77 containers of services. This level of detail has been

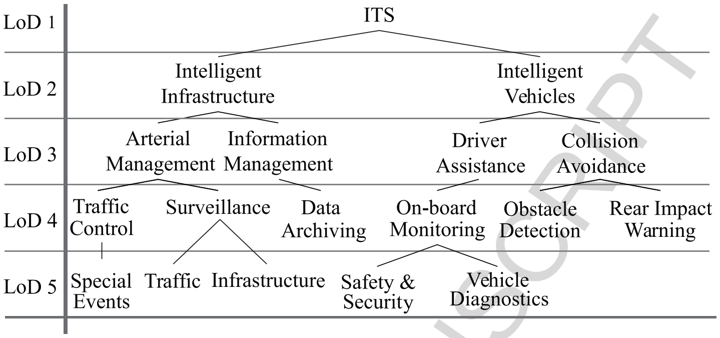


Figure 2: Part of RITA U.S. DoT taxonomy (until LoD 5).

chosen because it is an intermediate point between previous too generic and subsequent too detailed containers of services.

3.2. Selection of the Collection of Documents

The next step is the selection of the relevant information to apply semantic techniques. The 10 journals with the highest Impact Factors (IF) in the field of the ITS and, for each one, the 30 most important publications for the last 10 years (2005 to 2015) has been collected, giving as a result 300 publications, as shown in Table 1.

Notice that the selected information is grouped in collections of 30 papers. One of the drawbacks of using a collection of documents is that the weight of each keyword in each paper is different. One possible solution to overcome this issue is the IR feedback technique for relevance (Salton and Buckley, 1990). The main idea in this technique is that once certain retrieved documents have been considered as relevant or irrelevant by the user, the provided information is used to adapt the query so more relevant documents are retrieved in a subsequent search. However, the process of altering a query in the direction to relevant documents is an effective technique in information retrieval of an entire document, but not of specific parts of it. This paper proposes a novel method for the identification of paragraphs in the collection of documents as an alternative to the basic unit of analysis.

3.3. Discrimination of Paragraphs with Keywords

The main objective of the proposed Discrimination of Paragraphs with Keywords (DPK) is to retrieve only the most relevant information in the document collections in the form of paragraphs. The DPK is a method for selecting and discriminating paragraphs, filtering results according to the selected containers of services and providing the 150 most frequent words for each container of services. The main steps of the DPK method are illustrated in the Fig. 3.

Basically, the method stores only paragraphs that meet a search criteria (those paragraphs that contain exactly the keywordsToSearch()). The user is responsible for selecting the keywords, which are the previously considered containers of services. For each of the 77 containers of services, the DPK method analyzes the collection of documents and parses them, line by line, storing the related paragraphs in a temporary variable.

A paragraph is a component of the text that begins with a capital letter and ends with a full stop. All the words and characters are transformed to lowercase to facilitate the information management. When the search criteria of the DPK method is not satisfied, the entire paragraph is discriminated and discarded. Once the set of paragraphs related to a certain container of services is extracted and stored, they are sent to the function removing the stopwords, who must return the plain text. The last step of the DPK method is to find the absolute frequency of the 150 most frequent words within the

Table 1: Selected collection of documents.

	The most important publications	
ID	CONSIDERED JOURNALS	No.
A	Intelligent Transport Systems, IEEE Proc.	30
B	Intelligent Transport Systems, IET	30
C	Intelligent Transportation Systems Magazine, IEEE	30
D	Intelligent Transportation Systems, IEEE Trans.	30
E	Vehicular Technology, IEEE Trans.	30
F	Accident Analysis & Prevention	30
G	European Paper of Operational Research	30
H	Transportation Research Part A - Policy and Practice	30
I	Transportation Research Part B - Methodological	30
J	Transportation Research Part C - Emerging Technologies	30
	TOTAL COLLECTION OF PAPERS	300

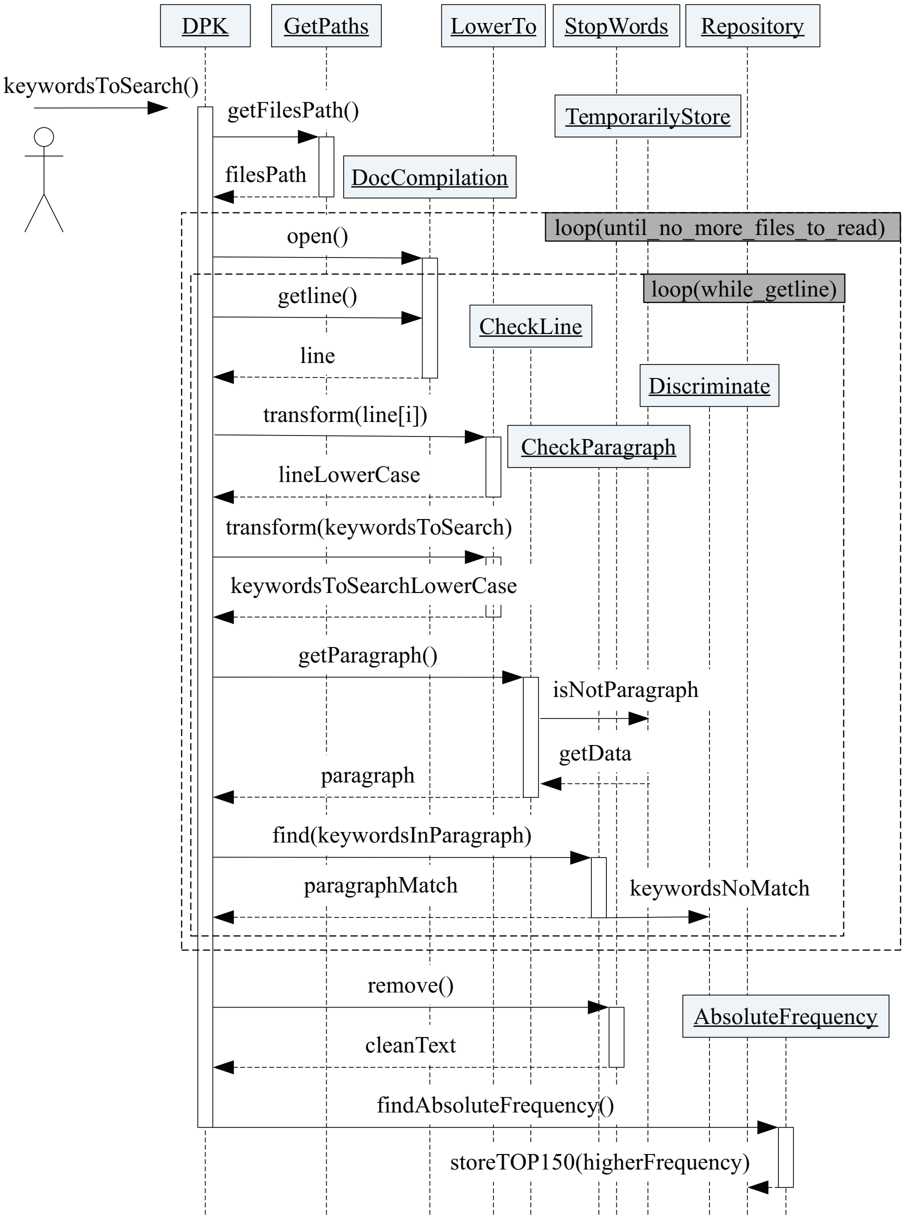


Figure 3: Sequence diagram of DPK method.

pieces of texts associated to the container of services, and store them in a repository, which is then used by the subsequent Information Retrieval with

Table 2: term/collection matrix 150×10.

Surveillance	Papers Collection of the Considered Journals										
Words	A	B	C	D	E	F	G	H	I	J	
Traffic	0.32	0.09	0.01	0.16	0.07	0.00	0.18	0.04	0.12	0.01	1
Surveillance	0.36	0.06	0.08	0.10	0.23	0.01	0.07	0.02	0.06	0.02	2
Time	0.41	0.05	0.02	0.03	0.18	0.00	0.06	0.00	0.25	0.00	3
Data	0.45	0.07	0.00	0.06	0.24	0.00	0.08	0.00	0.10	0.00	4
System	0.31	0.06	0.01	0.02	0.37	0.00	0.14	0.02	0.05	0.00	5
Vehicle	0.15	0.10	0.10	0.26	0.25	0.00	0.09	0.00	0.04	0.00	6
Real	0.38	0.08	0.05	0.03	0.29	0.00	0.06	0.00	0.11	0.00	7
Information	0.54	0.03	0.00	0.13	0.07	0.00	0.13	0.00	0.10	0.00	8
Estimation	0.63	0.08	0.00	0.00	0.00	0.00	0.12	0.00	0.18	0.00	9
Technology	0.17	0.07	0.07	0.05	0.62	0.00	0.02	0.00	0.00	0.00	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
word 150	…	…	…	…	…	…	…	…	…	…	150
	1	2	3	4	5	6	7	8	9	10	

Weighted Data in Paragraphs (IRWDP method).

3.4. Information Retrieval with Weighted Data in Paragraphs

The next step of the proposed methodology is the IRWDP method, which acts as a discriminating feedback system to obtain the weighted relative frequencies based on a specific search. Fig. 4 details how it is working. The IRWDP method work again with each of the container services of Fig. 2, and search their associated pieces of texts and 150 words in the stored repository by the DPK method. The aim of IRWDP is obtaining for each container of services the relative frequencies of the 150 words per collection of documents.

Table 2 shows an example of the obtained results in the case of "Surveillance" as the selected container of services. The first column lists its associated 150 words and the rest of the matrix is the relative frequency of these keywords in the selected paragraphs by the DPK method for this specific container.

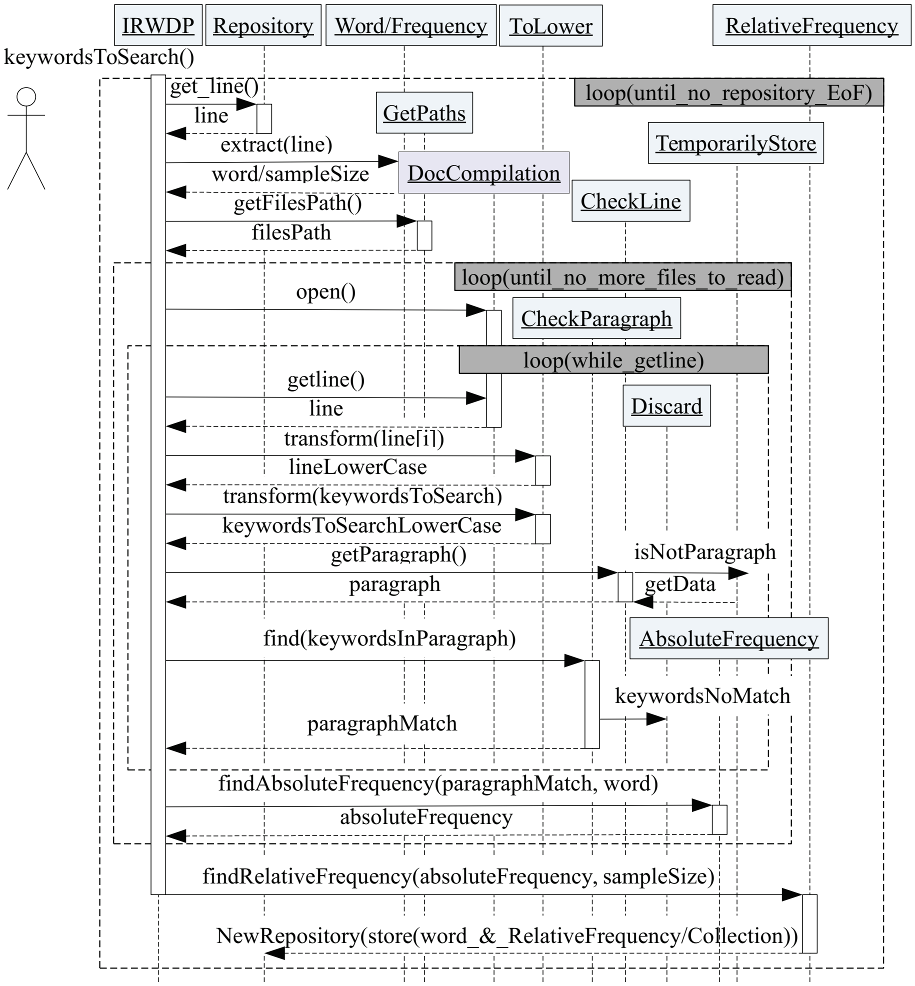


Figure 4: Sequence diagram of IRWDP method.

3.5. Knowledge Extraction

The previous obtained matrices for each container of services represent the extraction of unstructured knowledge. This is known as a vector space model, where information is summarized by column vectors in a term/collection of documents matrix. Mathematically, given a term/collection matrix $m \times n$ $A = (a_{ij})$ the nth term a_{ij} represents a weighted frequency term i in the collection j. The cosine of the rows of matrix A is a measure of the similarities among words. This value relies on the idea that the similarity of words depends on how many times they appear together, that is, their co-occurrence. However, the interpretation of the obtained results is difficult due to the high dimensionality of the model space. Nevertheless, one advantage of using the vector space model is that once a document collection is represented by columns in a high-dimensional space matrix, its algebraic structure can be exploited to reduce its dimensionality, always preserving the original vector space structure (Ye et al., 2004). The interpretation in this new space model is easier due to the reduction in the dimensionality of the space. Next methods apply different techniques to obtain a structured representation of knowledge.

3.6. Natural Language Processing (NLP)

LSI (Latent Semantic Indexing) is a well known semantic technique for building a semantic space (Deerwester et al., 1990), (Foltz, 1990a), (Foltz, 1990b), (Foltz, 1996). LSI, also known as LSA (Landauer et al., 1998), (Landauer and Dutnais, 1997) (Latent Semantic Analysis) is an indexing and retrieval method which examines the similarity of the contexts in which the words appear, creating a reduced dimension where the characteristics of the more similar words are those that are closer to each other. This technique is based on the principle that the words or terms that are used in the same context tend to have certain similarity. LSA is used to predict textual coherence, understanding, contextual disambiguation of homonyms and the generation of the inferred core meaning of a paragraph. LSA assumes that the dimensionality, wherein all relations of local context words are represented simultaneously, can be of great importance and reducing the dimensionality of the matrix of the observed data and the initial context number to one much smaller could produce better approximations to human cognitive relations (Landauer, 2002). There are different techniques and algorithms to reduce the matrix dimensionality. Some of the most popular are SVD (Singular Value Decomposition), PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), among others. In this paper we use the SVD technique, mainly because the focus of the paper is to propose a technique for create ontologies, and not to compare the dimensionality reduction techniques. LSA/LSI uses SVD as a method based on a linear algebra theorem (Leach, 1995), (Baker, 2005) to reduce the data matrix and to identify patterns in the relationship between the terms and concepts contained in a collection of unstructured text. It is no necessary to use any external dictionary, thesaurus or knowledge bases to determine these associations between words as they are derived from a numerical analysis of existing texts. SVD decomposes the rectangular matrix A into the product of three orthogonal matrices, an orthogonal matrix U, a diagonal matrix S and the transpose of an orthogonal matrix V. The theorem is usually expressed by:

$$A_{mn} = U_{mn} S_{nn} V_{nn}^{T}, \tag{1}$$

where U is an orthogonal matrix of $m \times n$ elements, $S_{nn} = diag(\sigma_1, \cdots, \sigma_n)$ is an $n \times n$ diagonal matrix containing the singular values σ_i of A, and V_{nn}^{T} is an $n \times n$ orthogonal matrix. SVD is closely related to standard eigenvalue-eigenvector decomposition of a square symmetric matrix. LSA uses the 150×10 term/collection matrix of Table 2, to construct the semantic space. In the matrix, each row corresponds to a single word in the corpus of publications and each column represents the collection of relevant documents in the Table 1.

3.7. Statistical Data Analysis

From the dimensionality reduction, can be drawn characterizations to predict or derive useful relationships between words, using clustering techniques and obtain a data structure of them.

***Step 1**. Application of Hierarchical Agglomerative Clustering*

Clustering algorithms allow to group a series of vectors according to a proximity criteria defined in terms of a given distance function. Generally the vectors of the same cluster share common properties. Using these groups, it can describe and build services within the multidimensional data set and express them as dendrograms or ultrametric trees, where pairs and triples of words can be visualized using simple and intuitive graphics.

Clustering algorithms have been applied to a large number of problems in a wide variety of research areas with the aim of identifying relevant distribution patterns that remain hidden. The hierarchical clustering builds a cluster hierarchy top-down (divisive) or bottom-up (agglomerative), by recursively splitting or merging clusters using some similarity metric. The split/merge process continues until a stopping criterion is met (i.e. number of clusters) (Sileshi and Gamback, 2009). The typical methods for hierarchical agglomerative clustering are: single-linkage, complete-linkage, average-linkage (Li et al., 2009) and the wards methods (Ding and He, 2002). The Ward and the average-linkage methods are the most popular ones (Everitt et al., 2011).

In this paper, the average-linkage algorithm has been chosen because of its robusticity (Everitt et al., 2011), its higher performance (Li et al., 2009) and the quality of provided clusters (Sileshi and Gamback, 2009).

In the average-linkage, the distance between two clusters is defined as the average distance between pairs of observations, one in each cluster. The average-linkage commonly joins clusters with small variations and tends slightly to produce clusters with the same variance. The hierarchical clustering results can be graphically represented as a tree-diagram or a dendrogram.

Step 2. *Application of the UPGMA Method to Build the Ultrametric Tree and to Extract Pairs/Triple Words*

One of the computational challenges of this study is to obtain a data structure to help in the final representation of an ontology. One solution proposed by (Gibas and Jambeck, 2001) was the implementation of phylogenetic trees. In computer science, there is a data structure that possesses the properties of phylogenetic trees called ultrametric trees.

The distance between two arbitraries vertices x and y of T, $disT(x, y)$, is the sum of the weights of the edges composing the path from x to y (Bockenhauer and Bongartz, 2007). Given a matrix of taxa (subjects or objects), two simple methods for building ultrametric trees can be used. The first one is called Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and the second is Weighted Pair Group Method with Arithmetic Mean (WPGMA). Both of them are agglomerative hierarchical methods using average-linkage technique. The UPGMA is widely used in bioinformatics to develop taxonomies with numerical data obtained from a set of taxa (Sokal and Sneath, 1963). This method constructs the bottom-up phylogenetic tree from the leaves (set of taxa). In the UPGMA method, distances are calculated using an arithmetic average depending on the number of elements in each cluster. Basically both methods, UPGMA and WPGMA, work in the same way. The only difference is the function of distance used in the last step. WPGMA makes use of the weighted average, which ensures that each taxon is equally participating in the final result. With the distance function used by WPGMA, each taxon contributes equally to the final result. UPGMA and WPGMA differ in the final result but not in the mathematical mechanism to achieve it. For the methodologies proposed in this paper, the UPGMA method is used because it is simpler, faster and have been widely used in the literature.

The taxonomy proposed by the U.S. DOT and RITA, Fig. 2, consists of two major groups, one of them focused on the Intelligent Infrastructures and the other one on Intelligent Vehicles. The total sample consisted of 34,738 paragraphs. In the case of Intelligent Infrastructure, the discriminated sample using the DPK method was of 1,519, discarding the rest of paragraphs because of their low relevance according to the considered containers of services.

Fig. 5 details the number of paragraphs associated to the different containers of services included under the general group of Intelligent Infrastructures. It can be noticed a clear trend of research on Traffic Control. These results can be clearly explained by the increasing investment of public authorities in the improvement of Road Infrastructure and Security. Fig. 6 details the same result but for the case of Intelligent Vehicles. A total of 843 paragraphs were discriminated from the initial sample using the proposed tools. Obtained research trends are more balanced among the different containers of services, but with more emphasis on Route Guidance. This is also a expected result since route guidance tools have become an important way of alleviating congestion in urban transport network and they are closely related to Traffic Control in the Intelligent Infrastructures.

Table 3 details the size reduction after applying DPK methods measured in paragraphs and MB. The reduction for IIwas 95.63% while the reduction in Intelligent Vehicles was 97.6%. After applying IRWDP method and the dimensionality reduction of LSA, it is possible to locate keywords of each

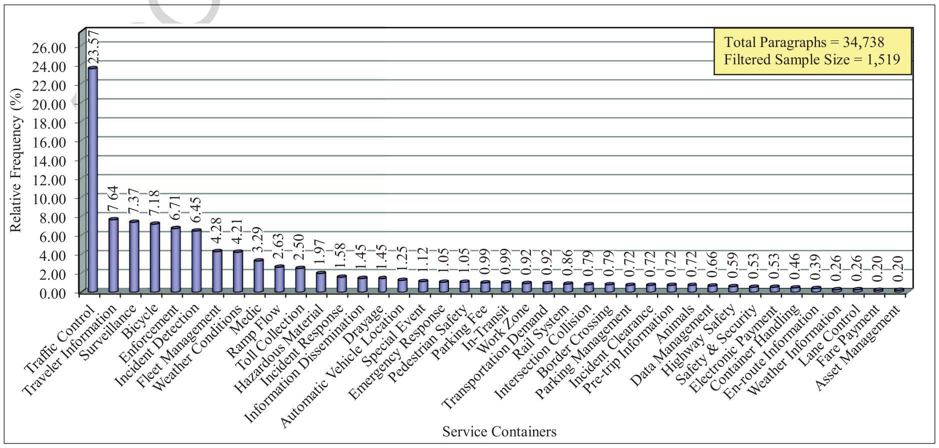


Figure 5: Relative Frequency of Paragraphs about Intelligent Infrastructures.

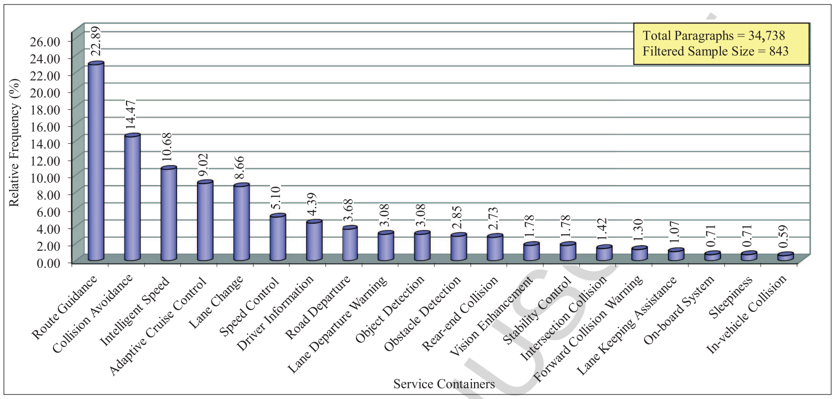


Figure 6: Relative Frequency of Paragraphs on Intelligent Vehicles.

container of services in a cartesian coordinate space. Table 4 shows a reduction to three dimensions for the particular case of "Surveillance" container using the predictive analytics tool, RapidMiner 5 (Rapid-I, 2012), (Lessmann et al., 2008).

The main problem of a three dimensional representation, is that results are more difficult to be interpreted. For this reason, it is preferable to consider only two dimensions and assume the loose of information in order to

Table 3: Results after applying DPK method.

	Sample size reduction after DPK		
ITS	TOTAL PARAGRAPHS	FILTERED PARAGRAPHS	SAMPLE SIZE REDUCTION
II	34 738	1 519	95.63%
IV	34 738	843	97.60%
	Useful Data Size after DPK		
	SIZE IN MB	REDUCED RATE	USEFUL DATA SIZE
II	13.2	95.63%	0.58 MB
IV	13.2	97.60%	0.58 MB

II: Intelligent Transportations, IV: Intelligent Vehicles.

Table 4: LSA 3D - Data Dimensionality Reduction.

ExampleSet (150 examples, 1 special attribute, 3 regular attributes)				
Row No.	Word	svd_1	svd_2	svd_3
1	Traffic	0.074	0.001	-0.042
2	Surveillance	0.077	0.037	0.009
3	Time	0.089	0.003	0.056
4	Data	0.093	0.025	0.026
5	System	0.077	0.079	0.021
6	Vehicle	0.073	-0.009	-0.066
7	Real	0.051	0.071	-0.031
⋮	⋮	⋮	⋮	⋮
150	word 150	⋯	⋯	⋯

benefit the interpretability of results. Table 5 and Fig. 7 detail the dimensionality reduction and the graphical representation considering only two dimensions for the same particular case of “Surveillance” container. In this graph, the diameter of each bubble represents the similarity between target words and color of these, represent each of the 150 terms. Then, the average-linkage model of agglomerative hierarchical clustering is applied to represent keywords of each container as a dendrogram or ultrametric tree using the UPGMA method. This way data is organized into subcategories that will be in turn divided in others until reaching the desired level of detail. The ultrametric distances are then those that meet the criteria of three points (the three-point condition) (Deonier et al., 2005) which say: d is an ultrametric tree in Q, if the elements in each three-element-subset of Q can be labeled by x, y, z such that:

$$d(x, y) \leq d(x, z) = d(y, z). \tag{2}$$

According to the ultrametric trees, pairs and triple of words have been extracted to build the ontology. Using pairs and triples of nearest words, it is possible to extract the information that later it is used to build the ontology. Fig. 9 shows the particular result for the case of “Surveillance” container. Following the same procedure with the rest of containers extracted from LoD4 taxonomy, the whole ontology is completed. Due to space limitations, it is not possible to include the complete ITS ontology or the tree diagrams.

Table 5: LSA 2D - Data Dimensionality Reduction.

ExampleSet (150 examples, 1 special attribute, 2 regular attributes)

Row No.	Word	svd_1	svd_2
1	Traffic	0.074	0.001
2	Surveillance	0.077	0.037
3	Time	0.089	0.003
4	Data	0.093	0.025
5	System	0.077	0.079
6	Vehicle	0.073	-0.009
7	Real	0.051	0.071
⋮	⋮	⋮	⋮
150	word 150	...	...

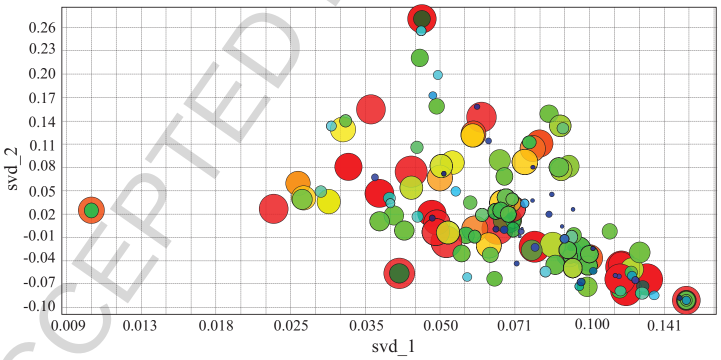


Figure 7: Plot Scatter 2D - Surveillance Container.

Using the open source ontology editor Protégé (Protégé, 2015), the developed ITS ontology can be modeled in OWL format or ported to others such as RDF, RDFS, etc.

4. RESULTS

In this section the results have been divided in two subsections. In the first one, it is a validation of the ontology built. In the second, it is conducted several experiments to evaluate the performance and scalability of

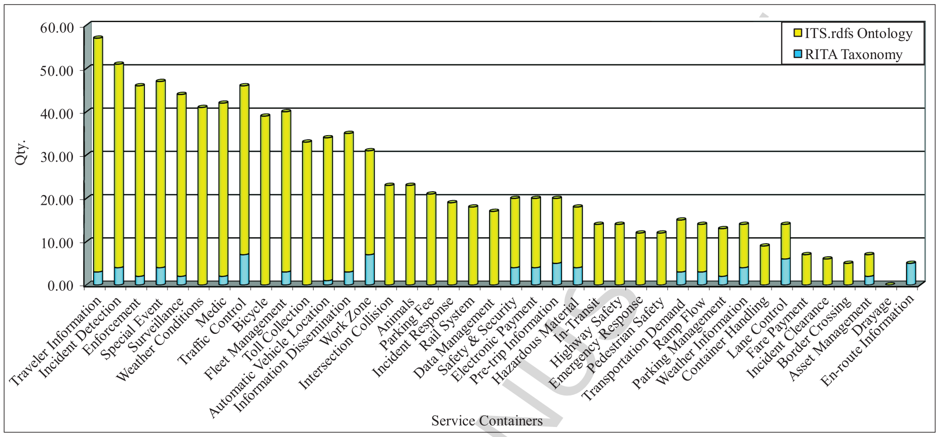


Figure 8: Containers of services in Intelligent Infrastructures.

the flow of information on embedded systems typically used in real urban and distributed environments.

4.1. Ontology validation

The taxonomy defined by the U.S. DOT and RITA addresses the classification of ITS applications. They provide a systematic organization of the

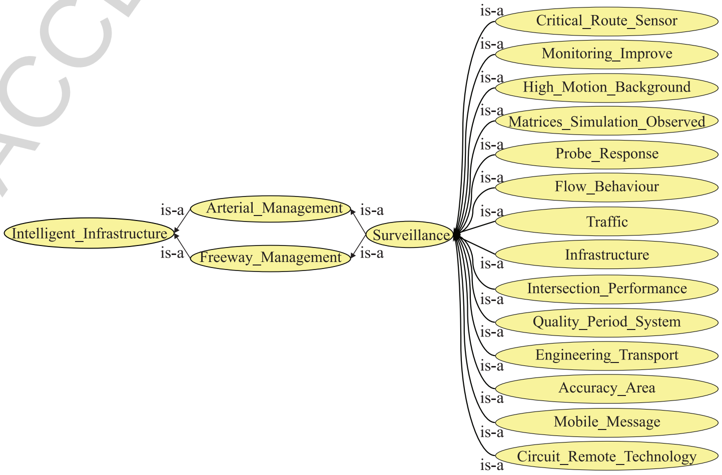


Figure 9: Part of ITS Ontology - Surveillance Container.

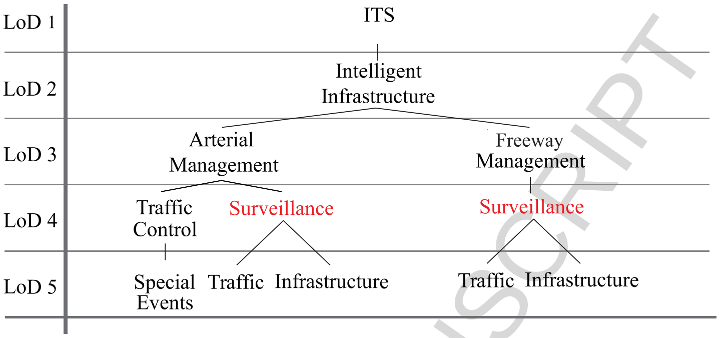


Figure 10: Homonymy in the RITA Taxonomy.

ITS field, giving names to groups of elements and final applications. A hierarchical structural model connects all terms in the taxonomy. Basically, this taxonomy considers two big categories: “Intelligent Infrastructures” with 14 applications and “Intelligent Vehicles” with 3 applications. Each of these 17 applications is divided into sub-applications with a brief summary of their benefits and information related to the area of interest. However, the taxonomy is only a simple classification that offers the costs and benefits of each application, without any semantic or logical structure in the data exchange.

As a difference, the developed ontology “ITS.rdfs” adds a descriptive logic. The data and metadata are stored in repositories, which provide access to all information on ITS applications and services discovered. The ontology is able to cope with the problems that RITA taxonomy cannot solve, such homonymy. Fig. 10 shows an example of homonymy problem in the case of Surveillance, both sub-classes of Arterial Management and Freeway Management within the category of Intelligent Infrastructure. Any system seeking a Traffic service about Surveillance within the taxonomy would receive both services, because it would be unable to distinguish one of them. Although the taxonomy contributes to the semantics of a term in the vocabulary, they do not define attributes between concepts and thus may cause confusion and conflicts. As a difference, the ITS.rdfs ontology is richer in terms of relations between terms. These relations allow to express the information within the domain without the need of duplicating terms; avoiding homonyms.

Fig. 11 shows that in the proposed ontology, Traffic and Infrastructure can be service containers, applications or services, belonging to the Surveillance class. Similarly, Surveillance is a sub-class of Arterial_Management

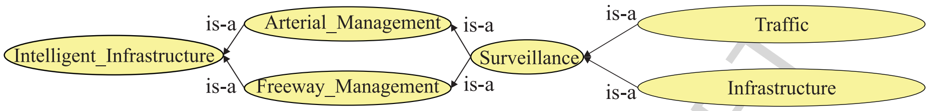


Figure 11: Homonymy solution in the ITS.rdfs ontology.

as well as Freeway_Management and these are themselves sub-classes of Intelligent_Infrastructure. As a difference to the taxonomy case, here there is not homonymy because they have different meanings and the nodes are in different semantic spaces. Thanks to namespaces, it is possible to avoid ambiguities in the result. Next figures compare the quantity (Qty.) of containers/services proposed by the U.S. DOT and RITA with the one obtained by the developed ITS.rdfs ontology, for Intelligent Infrastructures, Fig. 8 and for Intelligent Vehicles, Fig. 12. The quantity of services discovered about Intelligent Infrastructures in the developed ontology is 866, while the RITA taxonomy offers a maximum of 144 services/applications. In the case of Intelligent Vehicles, the total amount of discovered services is 449, against the 6 services/applications offered by the RITA taxonomy. The discovered services may be used as the basis for developing new applications/services in the field of ITS. The ontology developed will serve as a reference tool for information acquisition and construction of knowledge base systems that provide consistency, reliability and accuracy when retrieving information. The ITS.rdfs ontology enable sharing the knowledge and enable the collaborative work to function as common medium of knowledge between different actors involved in a urban, metropolitan and rural infrastructure.

4.2. Ontology Implementation

The created ontology ITS.rdfs with the proposed methods is used as a descriptive semantic service container, and the flow of information is treated as triplets SPO (Subject, Predicate, Object) for an Semantic Service (Gregor et al., 2012) (Semantic Communication Service ontology-based) capable of managing the flow of client/server requests in distributed urban environments. An important measure to check the performance is the throughput method as follows:

$$TputkB = \left(\frac{size(kB)}{RTT(sec.)}\right), \qquad (3)$$

where *RTT* is the "Round Trip Time" in seconds. However, this measure is not useful because the ontological data are expressed in triplets. Thus, the previous metric can be extended as follows:

$$TputkT = \left(\frac{no.triples/1000}{RTT(sec.)}\right), \tag{4}$$

which represents the total calculation on kiloTriplets of the ontology, over the *RTT* in seconds. To check, the overall performance has been tested storing the obtained ontology in a Berkeley DataBase (Oracle, 2014) on a PC-AMD Athlon (TM) 1200 MHz. The Semantic Service is capable of providing the communication support on distributed environments in conjunction with a set of base libraries like Redland (D., 2011c) (RDF Language Bindings) to interact with ontologies written in RDF and RDFS formats. A Raptor parser (D., 2011a) (RDF Syntax Library) is used to analyze the sequences of symbols, determine the grammatical structure and as a query language, Rasqal (D., 2011b) and (RDF Query Library) to build and run queries. Both, Rasqal and Raptor are designed to work with the Redland library. The goal of the distributed communication technology used in these tests is to manage the ontological information and interoperate with services

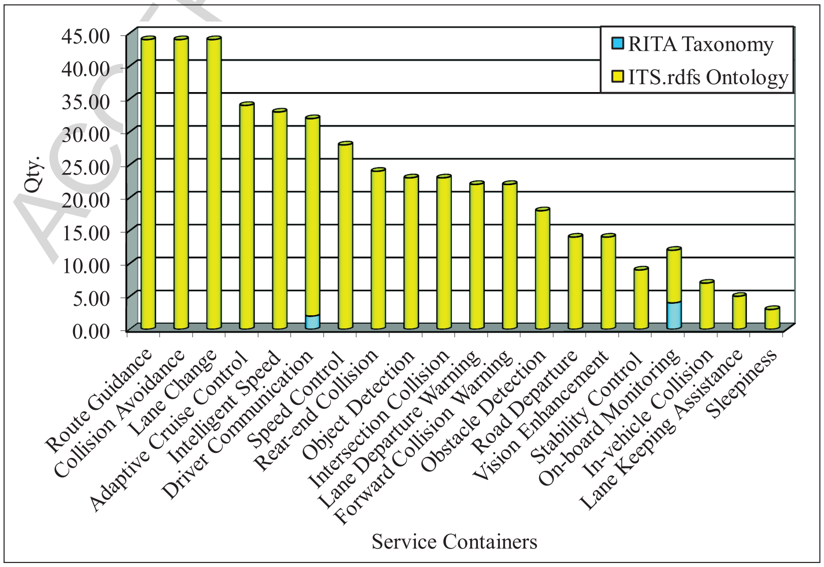


Figure 12: Containers of services in Intelligent Vehicles.

Table 6: Performance of the experiments.

Performance of the experiment 1	
Parsing and Storing the ITS.rdfs scheme in Berkeley DataBase	
$TputkT$ Average	3.57 kT/sec.
Total Time	806 ms.
Total Data Size	625.58 kBytes.
Performance of the experiment 2	
Read and analyze the temporal Model received from the Server	
Transfer Rate ($TputkB$)	176.40 kB/sec.
Total Delay	16 ms.
Total Data Size	2.84 kBytes.
Performance of the experiment 3	
Delay resolving the Query and building the response to sent to Client	
Total Delay	26 ms.
Total Data Size to Send	1.05 kBytes.
Total Delay (Client-Side)	41 ms.

discovered by the proposed methodologies in the previous sections. First, it is measured the performance parsing and storing the main ITS.rdfs scheme in a DataBase hosted on the PC-AMD. The performance was quite stable during the experiment, Table 6.

In urban environments, the different traffic services are mostly implemented in embedded devices. The next step was to estimate the system performance by adding new statements of a service in the stored ontology, Table 6. This service operates and runs on a device with ARM926EJ-S platform and the main function is to export the information that should be added to the ITS.rdfs ontology. The server (exporter) creates a RDF file that contains 14 statements (triples). This RDF is marshalled in a string and contains all the information that will be useful, in a client/server distributed environment, so that the client can access it. With the new 14 statements added, the ITS.rdfs ontology has now 2,896 triplets (added to the original 2,882

Algorithm 1 SPARQL query from the client (importer)

```
PREFIX kb: <http://edsplab.us.es/kb# >
CONSTRUCT {? Car_Counter kb: ior ?ior.
?x kb:serv_creator ?serv_creator }
WHERE{
?Car_Counter kb:ior ?ior.
?x kb: serv_creator ? serv_creator }
LIMIT 1
```

triplets). Moreover, the client (importer) that is deployed on another device with ARM926EJ-S platform performs a query to the SS. The SS manages the DataBase containing the ontology, looking for an outcome that satisfies a SPARQL (Members, 2015) request, Code 1. The delay resolving the query and building the response to be sent to the client is 26 ms. The weight of marshalling data that will be returned to the client is 1.05 kB, and this data contains all information in RDF format that the client needs to initiate the communication with the server in another device. The client obtain this response marshalled in 41 ms, Table 6.

In this section the obtained results from the analysis of information flow performance is presented. On one side was measured the performance parsing and stored the ITS.rdfs ontology in a Berkeley DataBase, obtaining an average $TputkT$ of 3.57 kT/sec. in 806 ms. It is measured the transfer speed reading and analyzing the received temporal model from the semantic server, reaching 176.40 kB/sec. with a delay of 16 ms. Finally, it is measured the delay resolving semantic query from the client, which reaches 26 ms. The SS marshall the result of the query and sends it to the client, this data packet weighs 1.05 kBytes. The client receives these results in 41 ms. The results demonstrate the feasibility of involving semantics implementation and the flexibility of representig metadata on ITS environments. Incorporating ontologies greatly favor the communication between devices and applications, achieving a common understanding between them.

5. Conclusion

This paper details a methodology for building ontologies considering IR and SLR as key steps. Technical proposals were applied in the construction of an ontology in the ITS domain and provides a guidance to develop ontologies in other areas. The field of study can be extended using as samples a greater amount of journals. Two new innovative methods have been proposed: DPK and IRWDP, which allows reducing the sample size of the study and assembling the terms/collection matrix with the relative frequencies of the words. The 150 most important words were taken on the outcome of the DPK technique. 150 words are more than enough for an average of 20 to 50 pairs/triples of words. LSA has some limitations in the order of the words as well as syntactical or logical relationships between them. The word order follows a consistent meaning for human understanding. To form pairs/triples of words, human intervention and control to maximize consistency was needed. The proposed methodologies, solve the major limitations, building ontologies from scientific articles, due to the lack of standardization for building ontologies. Using matrices, a dimensionality reduction, a hierarchical clustering and ultrametric trees were applied to extract pairs/triples of words for developing the ontology with the most relevant ITS services. With the addition of these techniques, it is possible to extract the information, collect, cut, sort and largely automate the creation of an ontology. One of the most important contributions of this paper is the proposed techniques for discriminating and classifying paragraphs according to their similarity. Using these techniques, between 95% and 98% of the total information was discarded. The discrimination of irrelevant data is extremely important for building ontologies to obtain subjects, predicates and objects with semantic meanings. Several performance tests were conducted on the obtained ontology using a Semantic Service, obtaining an average of 3.57 kT/sec. and 806 ms on a total weight of 625.58 kBytes, parsing and storing the full ontological scheme. Using a distributed client/server system implemented in an ARM9 embedded platforms, the performance of the data flow between them and the stored ontology in a Berkeley DataBase managed by the Semantic Service was measured. The experimental results demonstrate the feasibility and effectiveness of the approach. For future works, the proposed techniques could be extended, implementing different techniques and testing the feasibility of applying different algorithms to reduce dimensionality, clustering and also creating ultrametric trees, in order to optimize the results. It could be also considered a bigger size of samples and then make comparisons with the results obtained in this paper. In the IRWDP technique, the length of the words could be limited with a minimum of 4 characters and a maximum of 15 because sometimes, some unwanted characters may be part of the collection. Furthermore, it can be considered WPGMA method and then compare the results obtained in this research work with the UPGMA method.

6. References

Baker, K., 2005. Singular value decomposition tutorial.

Bockenhauer, H.-J., Bongartz, D., 2007. Algorithmic Aspects of Bioinformatics, 1st Edition. Springer-Verlag Berlin Heidelberg.

Chandrasegaran, S. K., Ramani, K., Sriram, R. D., Horvth, I., Bernard, A., Harik, R. F., Gao, W., 2013. The evolution, challenges, and future of knowledge representation in product design systems. Computer-Aided Design 45 (2), 204 – 228, solid and Physical Modeling 2012.

Chen, B., Cheng, H., June 2010. A review of the applications of agent technology in traffic and transportation systems. Intelligent Transportation Systems, IEEE Transactions on 11 (2), 485–497.

Chen, K.-H., Dow, C.-R., Guan, S.-J., Oct 2008. Nimbletransit: Public transportation transit planning using semantic service composition schemes. In: Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on. pp. 723–728.

D., B., 2011a. Raptor rdf parser library, 2011b.
URL http://librdf.org/raptor/libraptor.html

D., B., 2011b. Rasqal rdf query library, 2011c.
URL http://librdf.org/rasqal/librasqal.html

D., B., 2011c. Redland rdf library, 2011a.
URL http://librdf.org/docs/api/index.html

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., 1990. Indexing by latent semantic analysis. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 41 (6), 391–407.

Deonier, R. C., Tavar, S., Waterman, M. S., 2005. Computational Genome Analysis: An Introduction. Springer-Verlag Berlin Heidelberg.

Ding, C., He, X., 2002. Cluster merging and splitting in hierarchical clustering algorithms. In: Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. pp. 139–146.

Everitt, B. S., Landau, S., Leese, M., Stahl, D., 2011. Index. John Wiley Sons, Ltd, pp. 321–330.
URL http://dx.doi.org/10.1002/9780470977811.index

Fernandez, A., Ossowski, S., June 2011. A multiagent approach to the dynamic enactment of semantic transportation services. Intelligent Transportation Systems, IEEE Transactions on 12 (2), 333–342.

Fernandez, M., Gomez-Perez, A., Juristo, N., March 1997. Methontology: from ontological art towards ontological engineering. In: Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering. Stanford, USA, pp. 33–40.

Foltz, P., 1996. Latent semantic analysis for text-based research. Behavior Research Methods, Instruments, Computers 28 (2), 197–202.
URL http://dx.doi.org/10.3758/BF03204765

Foltz, P. W., Mar. 1990a. Using latent semantic indexing for information filtering. SIGOIS Bull. 11 (2-3), 40–47.
URL http://doi.acm.org/10.1145/91478.91486

Foltz, P. W., 1990b. Using latent semantic indexing for information filtering. In: Proceedings of the ACM SIGOIS and IEEE CS TC-OA Conference on Office Information Systems. COCS '90. ACM, New York, NY, USA, pp. 40–47.
URL http://doi.acm.org/10.1145/91474.91486

García, C., Padrn, G., Gil, P., Quesada-Arencibia, A., Alayn, F., Prez, R., 2012. Context model for ubiquitous information services of public transport. In: Bravo, J., Lpez-de Ipia, D., Moya, F. (Eds.), Ubiquitous Computing and Ambient Intelligence. Vol. 7656 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 350–358.
URL http://dx.doi.org/10.1007/978-3-642-35377-2_49

Gardels, K., June 1960. Automatic car controls for electronic highways. Tech. Rep. GMR-276, Research Lab. Warren, General Motors, Michigan.

Gibas, C., Jambeck, P., Apr. 2001. Developing Bioinformatics Computer Skills, 1st Edition. O'Reilly Media.

Gregor, D., Marn, S. L. T., Ariza, T., Barrero, F., 2012. An ontology-based semantic service for cooperative urban equipments. J. Network and Computer Applications 35 (6), 2037–2050.

Gruninger, M., Fox, M. S., 1995. Methodology for the design and evaluation of ontologies.

Guo, Z., Song, M., Wang, Q., Aug 2010. A framework of enterprise cloud application. In: Web Society (SWS), 2010 IEEE 2nd Symposium on. pp. 729–732.

Hall, T., Beecham, S., Bowes, D., Gray, D., Counsell, S., Nov 2012. A systematic literature review on fault prediction performance in software engineering. Software Engineering, IEEE Transactions on 38 (6), 1276–1304.

Kabbani, N., Tilley, S., Pearson, L., April 2010. Towards an evaluation framework for soa security testing tools. In: Systems Conference, 2010 4th Annual IEEE. pp. 438–443.

Landauer, T. K., 2002. On the computational basis of learning and cognition: Arguments from LSA 41, 43–84.

Landauer, T. K., Dutnais, S. T., 1997. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. PSYCHOLOGICAL REVIEW 104 (2), 211–240.

Landauer, T. K., Foltz, P. W., Laham, D., 1998. An introduction to latent semantic analysis. Discourse Processes (25), 259–284.

Leach, S., 1995. Singular value decomposition - a primer.

Lessmann, S., Baesens, B., Mues, C., Pietsch, S., July 2008. Benchmarking classification models for software defect prediction: A proposed framework and novel findings. Software Engineering, IEEE Transactions on 34 (4), 485–496.

Li, K., Wang, L., Hao, L., June 2009. Comparison of cluster ensembles methods based on hierarchical clustering. In: Computational Intelligence and

Natural Computing, 2009. CINC '09. International Conference on. Vol. 1. pp. 499–502.

Members, W., 2015. Sparql protocol for rdf", w3c recommendation.
URL http://www.w3.org/TR/rdf-sparql-protocol/

Mitropoulos, G., Karanasiou, I., Hinsberger, A., Aguado-Agelet, F., Wieker, H., Hilt, H.-J., Mammar, S., Noecker, G., Sept 2010. Wireless local danger warning: Cooperative foresighted driving using intervehicle communication. Intelligent Transportation Systems, IEEE Transactions on 11 (3), 539–553.

Oracle, 2014. Berkeley db java edition, 12c release 1 (library 12.1.6.0, version 6.2.31).
URL https://docs.oracle.com/cd/E17277_02/html/

Phan, K. A., Tari, Z., Bertok, P., April 2008. Similarity-based soap multicast protocol to reduce bandwith and latency in web services. Services Computing, IEEE Transactions on 1 (2), 88–103.

Protégé, 2015. Free and open source ontology editor and knowledge-base framework.
URL http://protege.stanford.edu/

Qu, F., Wang, F.-Y., Yang, L., November 2010. Intelligent transportation spaces: vehicles, traffic, communications, and beyond. Communications Magazine, IEEE 48 (11), 136–142.

Rapid-I, 2012. Interactive design.products: Rapidminer.
URL http://rapid-i.com

RITA, 2015.
URL http://www.its.dot.gov/index.htm

RITA U.S. DoT, U., 2015. Taxonomy of intelligent transportation systems application.
URL http://www.itslessons.its.dot.gov

Rockl, M., Robertson, P., May 2010. Data dissemination in cooperative its from an information-centric perspective. In: Communications (ICC), 2010 IEEE International Conference on. pp. 1–6.

Salton, G., Buckley, C., 1990. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science 41, 288–297.

Sileshi, M., Gamback, B., March 2009. Evaluating clustering algorithms: Cluster quality and feature selection in content-based image clustering. In: Computer Science and Information Engineering, 2009 WRI World Congress on. Vol. 6. pp. 435–441.

Soares, V., Farahmand, F., Rodrigues, J., July 2009. A layered architecture for vehicular delay-tolerant networks. In: Computers and Communications, 2009. ISCC 2009. IEEE Symposium on. pp. 122–127.

Sokal, R., Sneath, P., 1963. Principles of Numerical Taxonomy. Books in biology. W. H. Freeman.
URL http://books.google.com.py/books?id=3Y4aAAAAMAAJ

Tekli, J., Damiani, E., Chbeir, R., Gianini, G., Third 2012. Soap processing performance and enhancement. Services Computing, IEEE Transactions on 5 (3), 387–403.

Terziyan, V., Kaykova, O., Zhovtobryukh, D., May 2010. Ubiroad: Semantic middleware for context-aware smart road environments. In: Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on. pp. 295–302.

Thomas, T., van Berkum, E., June 2009. Detection of incidents and events in urban networks. Intelligent Transport Systems, IET 3 (2), 198–205.

Toral, S., Torres, M., Barrero, F., Arahal, M., September 2010. Current paradigms in intelligent transportation systems. Intelligent Transport Systems, IET 4 (3), 201–211.

USDOT, 2015.
URL http://www.its.dot.gov/index.htm

Wang, F.-Y., Zeng, D., Yang, L., Oct 2006. Smart cars on smart roads: An ieee intelligent transportation systems society update. Pervasive Computing, IEEE 5 (4), 68–69.

Ye, J., Janardan, R., Park, C. H., Park, H., Aug 2004. An optimization criterion for generalized discriminant analysis on undersampled problems. Pattern Analysis and Machine Intelligence, IEEE Transactions on 26 (8), 982–994.

Zhai, J., Cao, Y., Chen, Y., Oct 2008. Semantic information retrieval based on fuzzy ontology for intelligent transportation systems. In: Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on. pp. 2321–2326.

Zhang, H., Babar, M. A., Tell, P., 2011. Identifying relevant studies in software engineering. Information and Software Technology 53 (6), 625 – 637, special Section: Best papers from the {APSEC}.

Derlis Gregor was born in Asuncion, Paraguay, in 1980. He received the Bachelor Degree in Systems Analysis and the Computer Engineering from the American University, Asuncion, Paraguay, in 2007. Received the M.Sc. and Ph.D. Degrees in Electronic, Signal Processing and Communications from the University of Seville, Spain, in 2009 and 2013, respectively. He is currently Head of the Laboratory of Distributed Systems, Engineering Faculty of the National University of Asuncion (FIUNA), Paraguay. His research interest focuses on the application of Intelligent Transportation Systems (ITS). Interoperability in Sensor Networks, Embedded Systems and Instrumentation Systems.

Sergio Toral received the M.Sc. and Ph.D. degrees in electrical and electronic engineering from the University of Seville, Seville, Spain, in 1995 and 1999, respectively. He is currently a Full Professor with the Department of Electronic Engineering, University of Seville. His recent research interests include sensor networks and intelligent transport systems. Prof. Toral was a recipient of the Best Paper Awards from the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS in 2009 and *Institution of Engineering and Technology Electric Power Applications* in 2010-2011.

Teresa Ariza was born in Cadiz, Spain, in 1968. She received the M.S. and Ph.D. degrees in Computing Science from the University of Seville, Spain, in 1991 and 2000, respectively. She is currently a full Professor with the Department of Telematic Engineering, US. Her main research interests include real-time and distributed systems, middlewares, intelligent transportation systems, embedded operating systems and health applications.

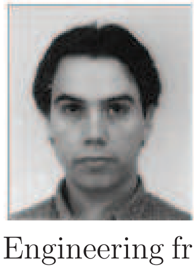
Federico Barrero received the M.Sc. and Ph.D. degrees in electrical and electronic engineering from the University of Seville, Seville, Spain, in 1992 and 1998, respectively.

In 1992, he joined the Electronic Engineering Department, University of Seville, where he is currently an Associate Professor. His recent interests include sensor networks and control of multiphase ac drives.

Dr. Barrero was a recipient of Best Paper Awards from the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS in 2009 and *IET Electric Power Applications* in 2010-2011.

Raúl Gregor was born in Asuncion, Paraguay, in 1979. He received the M.Sc. and Ph.D. degrees from the University of Seville, Spain, in 2006 and 2010 respectively. He joined Faculty of Engineering of the National University of Asuncion, Paraguay, in February 2009. Since 2012, he is the Head of the Laboratory of Power and Control Systems, Engineering Faculty of the National University of Asuncion (FIUNA). Prof. Gregor received the Best Paper Award from the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS in 2009, and the Best Paper Award from the Institution of Engineering and Technology ELECTRIC POWER APPLICATIONS, in 2012.

Jorge Rodas was born in Asuncion (Paraguay) in 1984. He received the Electronics Engineer Degree from the Engineering Faculty of National University of Asuncion in 2009. He received his Master's degree in 2012 in Signal Processing Applications for Communications from the University of Vigo (Spain). Since september 2011 he is with the Laboratory of Power and Control System, Engineering Faculty of National University of Asuncion. In 2013 he obtained a Master's degree in Electronics, Signal Processing and Communications from the University of Seville (Spain). He is a recipient of the Fundación Carolina Postgraduate Scholarship Award for his PhD study.



Mario Arzamendia received his bachelor degree in Electrical Engineering from the University of Brasilia (Brazil) in 2002 and his master degree in Electronic Engineering from Mie University (Japan) in 2009. From 2009 until 2013 he worked as a project leader at the Automation and Control Innovation Center (CIAC) of the Itaipu Technological Park. In 2013 he joined the Faculty of Engineering of the National University of Asuncion as a researcher and since 2014 he is coordinator of the Laboratory of Distributed Systems. His research interests include embedded systems and wireless sensor networks.

Highlights

1. We propose a methodology to build ontology's in the domain of ITS (Intelligent Transportation Systems) considering IR (Information Recovery) and SLR (Systematic Literature Review).
2. Two new methods have been proposed: DPK (Discrimination of Paragraphs with Keywords) and IRWDP (Retrieval with Weighted Data in Paragraphs).
3. The methods proposed allow reducing the sample size of the study.
4. Much information irrelevant has been discarded, achieving greater performance in the ontology construction.
5. The methodologies proposed can be used to build ontologies in any domains.