Expert Systems With Applications xxx (xxxx) xxx

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A semantic approach for document classification using deep neural networks and multimedia knowledge graph

Antonio M. Rinaldi *, Cristiano Russo, Cristian Tommasino

Department of Electrical Engineering and Information Technology, University of Napoli Federico II, 80125 Via Claudio, 21, Napoli, Italy

ARTICLE INFO

Keywords:
Multimedia topic detection
Document classification
Semantic analysis
Ontologies
Big data
Deep neural networks
Knowledge graph

ABSTRACT

The amount of available multimedia data in different formats and from different sources increases everyday. From an information retrieval point of view, this high volume and heterogeneity of data involves several issues to be addressed related to information overload and lacks of well structured information. Even if modern information retrieval systems offer to the user manifold search options, it is still hard to find systems with optimal performances in the document seeking process starting from a given topic. In recent years, several frameworks have been proposed and developed to support this task based on different models and techniques. In this paper we propose a semantic approach to document classification using both textual and visual topic detection techniques based on deep neural networks and multimedia knowledge graph. A semantic multimedia knowledge base has been exploited and several experimental results show the effectiveness of our proposed approach.

1. Introduction

The advent of the Internet, social networks, IoT and mobile technologies leads to a daily production of a huge number of data. This amount is constantly increasing and it represents rich informative contents if analyzed and related to other information.

Several companies from different business contexts, for example, quickly realized that having large collections of connected available data can be a gold mine for data analytic both from an operational and strategic point of view. On the other hand, academic researchers found that novel techniques based on big data and artificial intelligence give extraordinary results in many tasks with large volume of data (Rinaldi & Russo, 2018b, 2018c). As a matter of fact, this amount of data is a fundamental resource but some considerations have to be done regarding the possible side effects deriving from it. Some of the issues that need to be addressed are how to handle this huge amount of data, how to reliably categorize this data in order to be effectively reused and how to avoid the annoying problem of information overloading.

Among the various applications that can be realized starting from data categorization, there is the classification of documents. The task of classifying a document has the objective of assigning one or more classes to the analyzed document, making easier to manage them in a document collection. The techniques used to classify a document have been widely applied to different contexts, such as *genre classification* (Santini & Rosso, 2008), *sentiment analysis* (Melville, Gryc, & Lawrence, 2009; Tan & Zhang, 2008), *spam filtering* (Bíró, Szabó, & Benczúr, 2008), *language identification* (Li, Ma, & Lee, 2006; Takçı & Güngör, 2012), etc. Such techniques are often based on *semantic analysis*, focusing on the semantic relationships between terms in the analyzed document and the concepts that these terms represent. Semantic analysis methods are often based on the concept of *semantic network* (Sowa, 1987).

According to Woods (1988), the limit of semantic networks resided in the lack of a comprehensive and satisfactory theoretical definition. This issue lays in the vagueness of the conceptual role of a semantic network. The same indeterminacy, however, pushed the researchers to propose several techniques that today are able to implement efficient natural language analysis extracting useful information for different information retrieval tasks (Rinaldi, 2009; Rinaldi & Russo, 2018c). *Topic modeling* (Papadimitriou, Raghavan, Tamaki, & Vempala, 2000) is an approach based on the extraction of co-occurring lexical clusters in a documentary collection. Nowadays, it is one of the most innovative and widespread analytical methods. It is a set of unsupervised text mining techniques based on probabilistic approaches. Topic modeling aims to find patterns and structures in document collections. Initially, topic models were conceived to ease the task of browsing large document collections (Salton, Wong, & Yang, 1975). Most prominent methods used for topic modeling have been proposed over the years. The first is Latent Semantic Indexing (LSA) (Landauer, Foltz, & Laham, 1998), which extracts the underlying topics from a term-document matrix

* Corresponding author.
E-mail addresses: antoniomaria.rinaldi@unina.it (A.M. Rinaldi), cristiano.russo@unina.it (C. Russo), cristian.tommasino@unina.it (C. Tommasino).

https://doi.org/10.1016/j.eswa.2020.114320
Received 6 July 2020; Received in revised form 23 October 2020; Accepted 13 November 2020
Available online 18 November 2020

by applying singular value decomposition (SVD). The disadvantage of LSA is that it makes the orthogonality assumption among topics, which is in contradiction with human intuition about them. Hofmann (1999) addresses this issue by proposing a probabilistic method called LSA (pLSA): it models topics as word distributions. This model type has been improved by Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). LDA is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

In recent years, semantic networks have been often associated to ontologies (Caldarola & Rinaldi, 2016; Russo, Madani, & Rinaldi, 2020; Sarica, Luo, & Wood, 2020), which have become a silver bullet to represent semantics of data and allow machine-readable and machine-processable processes. While many definitions of ontologies have been provided over years, we will refer to an ontology as a formal, shared and explicit representation of a conceptualization of a domain of interest (Gruber, 1993). The natural duality between ontologies and graphs allows us to naturally represent the knowledge described in ontologies through graphs and vice versa. The use of graphs allows us to describe complex domains composed of highly connected nodes and to define efficient analysis metrics. Knowledge Graphs take large amounts of data from various data silos and try to add value to them so that it is possible to interconnect and re-use them in a more intelligent way (Paulheim, 2017).

The automatic use of topic detection and categorization allows quick access to document collections. Sometimes a disadvantage of many topic modeling methods (Hofmann, 1999; Landauer et al., 1998) is that they treat the categorization structure without considering the relationships between categories. Following subsumption theories (Brooks, 1987; Chandiok, Prakash, Siddiqi, & Chaturvedi, 2020), in our vision hierarchical or taxonomic structures are the preferred way in which concepts, subjects or categories are arranged in practice (Rinaldi, 2009) and the proposed framework aims to exploit such relationships for topic modeling.

In this work we propose a novel methodology that combines statistical information, NLP and deep learning techniques and technologies by providing a framework to categorize multimedia web documents using semantic analysis, ontologies and metrics based on *multimedia semantic similarity*. The additional information is extracted from a general knowledge base. The semantic knowledge base follows the work presented in Rinaldi and Russo (2018b). A semantic topic detection method is also proposed, which is the result of a combined textual and visual analysis of the original document.

The paper is organized as follow: Section 2 explores the literature and the state of the art regarding Topic Modeling and Detection techniques and technologies; Section 3 the proposed approach is presented and the whole architecture of the developed system is showed together with the proposed textual and visual classification methodology; Section 4 is devoted to the presentation and discussion of the experimental strategy and results; eventually, conclusions and future research directions are in Section 5.

2. Related works

In this section we explore the literature related to textual and visual topic detection. Approaches for textual topic detection aim to find keywords from a text document in order to build a set of representative terms for the given document. Several methodologies have been proposed over the years and they can be grouped according to the used approaches. In the statistic approach, the goal is to identify the relevance of a term according to statistical properties, such as TF-IDF (Sparck Jones, 1972), N-Grams (Cavnar, Trenkle, et al., 1994), etc. The linguistic approach uses *features* based on linguistic, syntactic, semantic, lexical properties and similarity functions to extract representative keywords. The third class makes use of machine learning techniques, e.g. Naive Bayes (Zhang, 2004), Support Vector Machine (Suykens & Vandewalle, 1999), etc. The keyword extraction is the result of a trained model able to predict significant keywords. Other approaches are based on combinations of the above-cited groups. Other parameters such as *word position*, *layout feature*, HTML tags, etc. can also be used.

According to the previous classification, the author in Hulth (2003) makes use a combination of both machine learning and linguistic /semantic approaches. The extraction of keywords is based on NP-chunks and POS tags which gives a better precision in results instead of N-grams. In Matsuo and Ishizuka (2004) co-occurrence is exploited for the extraction of keywords from a single document. Different information related to the vector space model and genetic algorithms are used in Alguliev and Aliguliyev (2005) to measure the score between the sentences and the weights associated to features. A similar approach is based on the unsupervised technique called MCMR (Maximum Coverage and Minimum Redundant) which allows to extract more relevant and non redundant sentences from the original document. In Hu and Wu (2006) the authors use linguistic features to represent the terms relevance, also based on the position of a term in the document. The authors of Xu, Yang, and Lau (2010) have proposed new features for keywords extraction and the creation of a title for documents using *Wikipedia* as a knowledge source. A semantic graph model to represent documents is presented in Rinaldi and Russo (2018a). An iterative approach for keywords extraction based on relations between different granularities (words, sentences, topics) is presented in Wei (2012). In this approach, a graph containing relations between different nodes is built, then the score of each keyword is calculated through an iterative algorithm. Probabilistic models for topic extraction are analyzed in Alghamdi (2015). A special eye is put on previously cited models such as LSA, PLSA, LDA which are very useful algorithms in topic modeling of web documents. Their advantages and disadvantages have been treated in Section 1. Several works (Jelodar et al., 2019; Prabhakar Kaila, Prasad, et al., 2020) are based on LDA algorithm and its variants, mainly differing for their application fields and used techniques. As an example of this category, Khalid and Wade (2020) makes use of topic detection from conversational dialogue corpus. Parallel Latent Dirichlet Allocation (PLDA) Model is employed by clustering a vocabulary of known similar words based on TF-IDF scores and Bag of Words (BOW) technique.

Visual topic detection makes use of algorithms for the extraction of several visual features (global features, local features, deep features). The combined use of this kind of features has been proven to be an effective strategy (Gavrila & Munder, 2007). Selecting the best features among the ones available, it is possible to obtain a better results with respect to the ones obtained using a single feature. In the early content-based information retrieval (CBIR) algorithms, global features have been widely used to characterize images based on color, shape and texture (Cao, Wang, Zhang, & Zhang, 2011; Wang & Hua, 2011). Subsequently, the task of *image retrieval* evolved also considering local features of images. Algorithms such as SIFT (Scale-Invariant Feature Transform) (Lowe, 1999) and SURF (Speeded Up Robust Features) (Bay, Ess, Tuytelaars, & Gool, 2008) have been developed with the goal of extending the retrieval task introducing approaches based on *key-points matching*. With the advent of *deep neural networks*, the accuracy in the process of *large-scale image recognition* notably improved with respect to results obtained using traditional algorithms which make use of *hand-designed features*. The approaches based on *deep learning*, have also shown that they can be applied in very diverse contexts, from the analysis of *big data* to *computer vision*, *pattern recognition*, *natural language processing* and recommendation systems (Liu et al., 2016). Recent studies have shown how *deep learning* could be used with success in several application fields using image based approaches, from the detection of diseases affecting plants based on

visual features (Mohanty, Hughes, & Salath, 2016) to the detection of damages in civil structures using automatic approaches of deep features extraction (Lin, Nie, & Ma, 2017).

Few works focused on methods combining both textual and visual information from a semantic point of view. Li, Joo, Qi, and Zhu (2016) proposes a multimodal topic and-or graph (MT-AOG) to jointly represent textual and visual elements of news stories and their latent topic structures. It is a method for automatically detecting and tracking news topics from multimodal TV news data. BERT (Asgari-Chenaghlu, Feizi-Derakhshi, Balafar, Motamed, et al., 2020) is a more recent work where a graph mining technique is employed to enhance the resulting topics with aid of simple structural rules. Named entity recognition from multimodal data, image and text, labels the named entities with entity type and the extracted topics are tuned using them.

In this paper we propose a semantic approach for document classification. We make use of deep neural networks frameworks and a multimedia knowledge base as a ground model for semantic connections. Moreover, we show that by combining textual and visual topic detection it is possible to improve the performances of the proposed method.

3. The proposed approach

In this section we describe all aspects of our approach for topic detection. Our methodology combines statistical information, NLP and deep learning techniques and technologies to categorize multimedia web documents using semantic analysis, ontologies and metrics based on semantic similarity, multimedia features and deep neural networks. The approach is implemented by means of a complete modular framework with a high degree of generalization using a general multimedia knowledge base containing textual and visual representations of concepts. Moreover, a combined textual and visual analysis of the original document is also proposed to define and implement a new semantic topic detection method. The rest of the section is devoted to a detailed description of the system architecture highlighting the main features of each module of the proposed framework. The multimedia knowledge base and the ontological model it relies on are also extensively described. Then, we describe in detail the proposed *Topic Detection* technique.

3.1. System architecture

The system architecture, shown in Fig. 1, is organized around different modules in charge of computing specific tasks. The first module is the pre-processing module; its task is to apply transformations to textual information and to store pre-processed text into document repository while images are stored in the image repository. Our system implementation is based on the document collection named *DMOZ* (DMOZ, 2002), an open multilingual web directory. Although the system uses the multimedia document collection in this directory, it could be used as well with different input sources, thanks to the nature of the implemented modules, which allow to interact with any existing web directory provided with RDF description (Miller, 1998). We put in evidence that the system could be used on different document collections by editing the pre-processing module. The system retrieves a list of URLs from DMOZ representing the documents contained in the web directory through the *RDF Fetcher*. The *Scraper* block takes care of making HTTP requests to the index-pages of the collected URLs and it sends the content of the response entry (HTML source) to the *DOM Parser*, which executes an analysis on the web page code in order to recognize significant data. The document is then downloaded splitting the text from the images. The images are transferred to the another macro-module (i.e. the *Visual Feature Extractor*) which is in charge of extracting multiple visual features from images. The textual part is instead subjected to a pre-processing phase in which cleaning operations are performed (*Text Document Pre-Processor* block in first macro module). The cleaning operations on the text are: (i) tags removing, (ii) stop words deleting, (iii) elimination of special characters, (iv) stemming. The third macro-module is the *Topic Detection* module, which uses algorithms based on the analysis of text and images to recognize the topic of the document, and the *Multimedia Knowledge Graph*, implemented in a NoSQL graph database. The proposed method involves a combined textual and visual topic detection, with an appropriate combining function; it is able to classify the main topic of the document. We propose a semantic textual analysis called *SEMREL* combined with visual descriptors. The output of the topic detection module is provided as input to the *Taxonomy Classificator* to generate, starting from a concept, a classification taxonomy with the help of our multimedia knowledge graph. The proposed metric and approach have been compared with baselines and the results are shown and discussed in the experimental Section 4.

3.2. The multimedia knowledge graph

The general knowledge base realized in this work is based on a multimedia ontology model proposed in Rinaldi (2014) implemented as a knowledge graph stored in a NoSQL database. The conceptual model representation uses *signs*, defined in Danesi and Perron (1999) as "*something that stands for something, for someone in some way*". A concept can be represented in various multimedia forms such as text, images, gestures, sounds and any way in which information can be communicated as a message. Each type of representation has properties that distinguish them. The structure of the model is composed of a triple $\langle S, P, C \rangle$ where S is the set of signs; P is the set of properties used to link signs with concepts; C is the set of constraints defined on the set P.

The approach used in this paper configures itself as multi-modal representation because we use two different types of input representations, text and images. According to the terminology used in the ontology model, they are our signs. The properties are semantic-linguistic relations and the constraints are given by validity rules applied to the properties with respect to the considered multimedia. Knowledge is represented by an ontology logically represented by a Multimedia Knowledge Graph. The latter can be seen as a graph where the nodes are concepts and the arcs are relations between them. It has been implemented using a NoSQL technology, in particular Neo4J graph DB (Webber, 2012). The population of the Knowledge base has been carried out by importing a limited number (up to 20) of multimedia representations (i.e.) per concept from *ImageNet* (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009), which is directly mapped with the lexicon-semantic dictionary *WordNet* (Miller, 1995). The concept is a set of multimedia data that represent an abstract idea at a high level of conceptualization. The language chosen to describe this model is the DL version of the *Web Ontology Language* (OWL) (McGuinness, Van Harmelen, et al., 2004), a markup language that offers maximum expressiveness while preserving completeness and computational decidability. It allows the declaration of disjoint classes to assert that a word can belong to a given syntactic category. Furthermore, it is possible to declare the unions of classes used to specify these domains, ranges and properties to relate concepts and multimedia nodes. The Fig. 2 shows the proposed ontological model.

The class *MM* with the respective sub-classes represents all the possible signs of our ontology, in this paper they are restricted to textual and visual case. The classes do not have common elements and therefore can be considered disjoint. In Table 1 we show some of the properties for the definition of domain and co-domain. The attributes of the *Concept* and *Multimedia* classes are also described. The concept has as attribute *Name* which represents the name of the concept and the field *Description* which contains a small description of it. The common attributes of the MM are *Name* and *ID*. Each subclass has its own set of features depending on the nature of the media. In the *Visual* case (i.e. images), global, local and deep features are used:

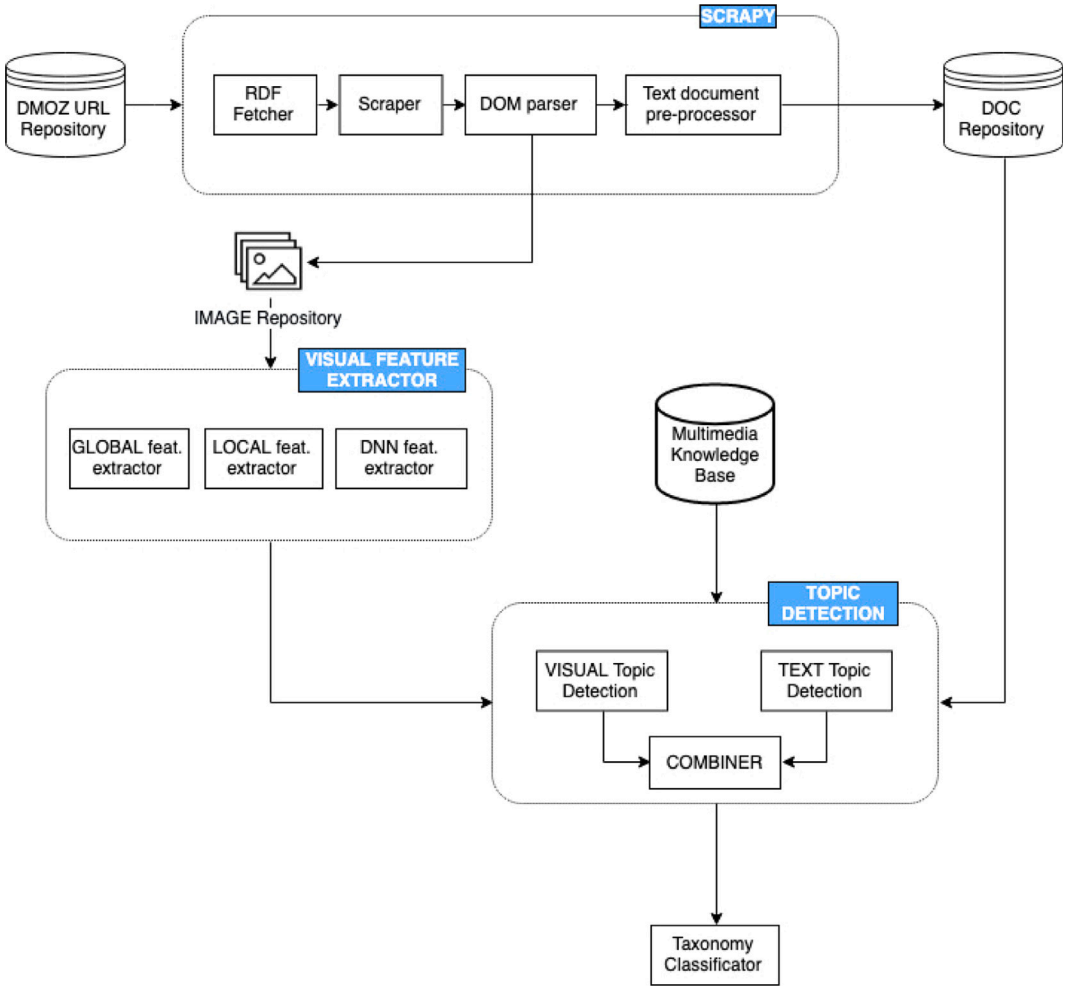

Fig. 1. System architecture.

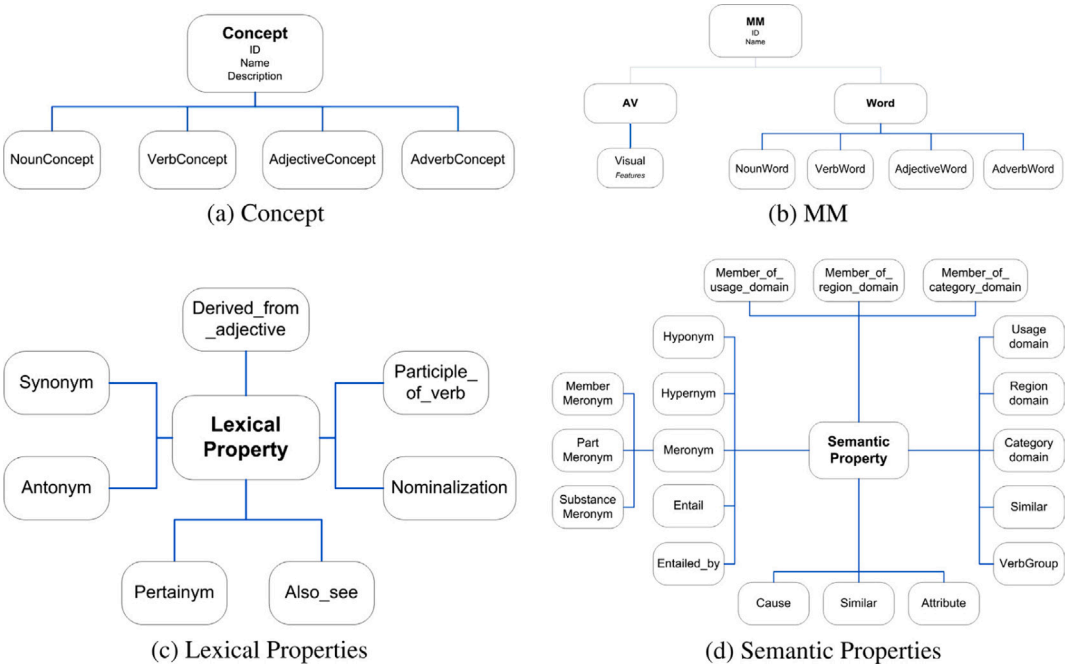

Fig. 2. Concept, multimedia and linguistic properties.

Table 1
Properties.

Property	Domain	Range
hasMM	Concept	MM
hasConcept	MM	Concept
hypernym	Nouns and verbs	Nouns and verbs
holonym	Noun	Noun
Entailment	Verb	Verb
Similar	Adjective	Adjective

- *Global Feature — Pyramid Histogram of Oriented Gradients* (*PHOG*) (Bosch, Zisserman, & Munoz, 2007): it is based on the occurrences of the gradient orientations in localized portions of an image. This method is similar to the *Edge orientation Histogram* and to the *Scale-Invariant Feature Transform* but it differs from them because it is calculated on a dense grid of uniformly spaced cells and uses a normalization of the local contrast overlapping to have a better accuracy.
- *Local Feature — Scale Invariant Feature Transform* (*SIFT*) (Lowe, 1999): it is a local feature that creates an invariant descriptor of translation, rotation, and scaling; robust in perspective and lighting variations. In its original formulation, the descriptor uses a method to recognize points of interest of a gray-scale image (key-points).
- *Deep Feature — CNN feature* (*VGG16*) (Simonyan & Zisserman, 2014): The deep feature used in this work was created using the output vector of the last *max pooling layer* (size: $7 \times 7 \times 512$) of *VGG16* (Gopalakrishnan, Khaitan, Choudhary, & Agrawal, 2017) (proposed by Visual Geometry Group of the University of Oxford). The strategy of using the output of the last pooling layer as a deep feature has been widely discussed and validated in a recent scientific article (Zhi, Duan, Wang, & Huang, 2016). The choice to use this architecture lies in the good results (performances) obtained on the ImageNet dataset (92.7% top-5 test accuracy — winner of the classification and localization task in the 2014 ILSVRC challenge).

Some extracted features, such as SIFT and VGG16, are natively presented as multi-dimensional arrays, hence we need to further process them. For example, in case of VGG16 features, a reshape of the vectors was computed in order to have a one-dimensional output. Moreover, a *base64* encoding was performed on visual descriptors due to the high dimension of the deep descriptor.

The links in the multimedia knowledge graph are represented as *ObjectProperties* and are defined with respect to constraints which depends on the syntactic category or property type (lexical or semantic) described in Table 1:

For example, the *hypernym* property is only applicable between noun–noun pairs or between verb–verb pairs. Each multimedia is related to the concept it represents through the ObjectPropety *hasConcept*, vice versa the property *hasMM* is used. These are the only properties that can be used to link concepts with multimedia, the other properties are used to link multimedia with multimedia or concepts to concepts.

As shown in Fig. 2, the semantic and lexical properties are also described as a hierarchy. The use of union classes simplifies the domain rules but at the same time does not exhibit perfect behavior. For example, the hypernym property, as the rule is defined in Table 1, allows the relationship between nouns and verbs. To solve this problem several constraints have been added, some of them are shown in Table 2:

In some cases, the existence of a property between two or more individuals leads to the existence of other properties. For example, if a concept A is a hyperonym of a concept B, the concept B is a hyponym of A. These characteristics are described by specifying property features. Some examples are shown in Table 3:

The model and the relative multimedia knowledge graph have been implemented in a *graph-db* based on the *property-graph-model* (Angles, 2018).

Table 2
Constraints for properties.

Constraint	Class	Property	Constraint range
AllValuesFrom	Noun	Hypernym	Noun
AllValuesFrom	Adjective	Attribute	Noun
AllValuesFrom	Verb	Also see	Verb
AllValuesFrom	Noun	Hyponym	Noun

Table 3
Features of properties.

Property	Features
hasMM	Inverse of hasConcept
hasConcept	Inverse of hasMM
hyponym	Inverse of hypernym: transitivity
hypernym	Inverse of hyponym: transitivity
verbGroup	Symmetry and transitivity

3.3. Topic detection

The *Topic Detection* task is carried out by different modules called: *Text Topic Detection*, *Visual Topic Detection* and *Combiner*.

The *Text Topic Detection* analyzes the textual content of a document to obtain a concept called *Topic Concept*, which represents the main topic of the document. In this paper we present a novel algorithm for text topic detection called SEMREL.

3.3.1. SEMREL

SEMREL is an algorithm based on the *bag-of-words* representation model. Starting from the cleaned document, the *tokenization* is performed to obtain the list of words contained in the document. The list is processed in the *Word Sense Disambiguation* step (see Section 3.3.1.1), during which the right meaning to the terms is assigned. Then, Dynamic Semantic Networks (DSNs) from the knowledge base (Section 3.2) are generated for all the terms, in the *SN Extractor* step. For each concept, the intersections between its generated semantic network and those of the other concepts are calculated by counting the number of nodes in common. The common nodes correspond to the degree of representation of the concept considered with respect to the entire document. This measure is called *Sense Coverage*. The latter factor would favor the more generic concepts and for this reason a scaling factor depending on the *depth* of the considered concept is used, calculated as the number of jumps to get to the root with the only hypernymy relationships. The *TopicConcept* is the one showing the best trade-off between the *SenseCoverage* and the *Depth*. Eq. (1) shows the formula used for calculating the topic concept of a given document.

$$TopicConcept = max(depth(C_i) * Coverage(C_i)) \quad (1)$$

where C_i corresponds to the ith concept resulting from the WSD step. Only concepts of type *noun* are considered from the WSD list, since we are looking for the topic of the document. In the Algorithm 1 we show, through pseudo-code, the logic used to find the topic concept.

We provide an example showing the flow of the topic detection process to help the reader in better understanding our approach. We consider the following textual document, whose topic is "beagle": "The beagle is a breed of small hound that is similar in appearance to the much larger foxhound. The beagle is a scent hound, developed primarily for hunting hare" The first step consists in cleaning the document performing a text normalization, i.e. removing stopwords, performing tokenisation, lemmatisation, etc. The output of this operation is a list of words, which will look like the following one: "[beagle, breed, small, hound, similar, appearance, much, large, foxhound, beagle, ...]". The second step is the above-cited Word sense disambiguation (WSD), further detailed in Section 3.3.1.1. With this operation, we are able to recognize semantic information, since each word is transformed into a WordNet synset (a synset is a set of synonyms) which would

Algorithm 1 Topic Concept Algorithm

```
procedure TopicConcept(ConceptList)
    BestConceptScore = 0
    for each concept C_i in ConceptList (after WSD) do
        ScoreC_i = 0
        DSN_C_i = BuildDSN(C_i)
        CoverC_i = 0
        for each concept C_j ≠ C_i in ConceptList do
            DSN_C_j = BuildDSN(C_j)
            NumberOfCommonConcept = Match(DSN_C_i, DSN_C_j)
            CoverC_i = CoverC_i + NumberOfCommonConcept
        end for
        ScoreC_i = depth(C_i) * CoverC_i
        if BestConceptScore < ScoreC_i then
            BestConceptScore = ScoreC_i
            TopicConcept = C_i
        end if
    end for
    return TopicConcept
end procedure
```

Table 4
Scores computation.

Concept	Score
hound.n.01	60
foxhound.n.01	32
beagle-n-01	**64**

represents a concept. The result would be similar to the following: "[beagle.n.01, breed.n.02, ..., hound.n.01, ...]". We explicit point out that we have used the WordNet notation (lemma.POS.senseNumber). The next step is the building of a semantic network for each synset previously recognized. Each semantic network will result in a list containing the hyponyms. Assuming that synsets with highest sense coverage are hound.n.01, foxhound.n.01 and beagle.n.01, with values 10, 9, 8, and depths 3, 4, 4 respectively. At this stage we also consider the occurrences, in case of hound.n.01 and beagle.n.01 is 2. The last step is the computation of Eq. (1), with the max value corresponding to the synset "beagle.n.01" as shown in Table 4.

3.3.1.1. Word sense disambiguation. The word sense disambiguation tries to mitigate the problem related to *polysemy* of terms. The task associates the correct meaning to terms by comparing each sense of a term with all the senses of the others. The similarity between terms is calculated through a linguistic-based approach and a metric computing their *semantic relatedness*, described in Rinaldi (2009), briefly recalled here. Weights σ_i are assigned to the properties of the ontological model described in Section 3.2. The goal is to discriminate the expressive power of relationships. The values of the weights are shown in Table 5 and follows the ones proposed in Castano, Ferrara, and Montanelli (2003) and Rinaldi (2009).

The metric is based on a combination of the best path between pairs of terms, and the depth of their Lowest Common Subsumer (LCS), expressed as the number of hops to get to the root of the tree with the only hypernymy relationships. The metric is normalized in the range $[0, 1]$ (1 when the length of the path is 0 and 0 when the length is infinite). The depth factor is used to give more weight to specific concepts (low level and therefore with high depth) than generic ones (low depth). About that a non-linear function is used to scale the contribution of the sub-ordinates of the upper level and increase those of a lower level. The best path is calculated as follows:

$$l(w_1, w_2) = min_j \sum_{i=1}^{h_j(w_1,w_2)} \frac{1}{\sigma_i} \tag{2}$$

Table 5
Property weights.

Property	Weight
Antonymy	0.8
Attribute	0.7
Category domain	1
Cause	0.6
Derived	0.8
Entailed by	0.7
Entailment	0.7
Hypernym	0.9
Hyponym	0.9
Member holonym	0.5
Member meronym	0.5
Member of category domain	1
Nominalization	0.7
Part holonym	0.7
Part meronym	0.7
Principle of	0.7
See also	0.6
Similar to	0.5
Substance holonym	0.5
Substance meronym	0.5
Synonymy	1

where l is the best path, w_j and w_i are the terms, $h_j(w_i, w_j)$ corresponds to the number of jumps of the jth path and σ_i corresponds to the weight of the ith edge of the jth path.

The Semantic Relatedness Grade of a document is then calculated as proposed in (Rinaldi, 2009):

$$SRG(v) = \sum_{(w_i,w_j)} e^{-\alpha \cdot l(w_i,w_j)} \frac{e^{\beta \cdot d(w_i,w_j)} - e^{-\beta \cdot d(w_i,w_j)}}{e^{\beta \cdot d(w_i,w_j)} + e^{-\beta \cdot d(w_i,w_j)}} \tag{3}$$

where (w_i, w_j) are pairs of terms in the document v, α and β are parameters whose values are set by experiments. The WSD process calculates the score for each sense of the term considered using the proposed metric. The best sense associated with a term is the one which maximizes the SRG obtained by the semantic relatedness between all the terms of the document. The variant of this process is related to the choice of a context window (*window of context*) reduced to parts of the document. For this purpose, the document is divided into grammatical periods, interspersed with linguistic punctuation points such as the dot, question mark and exclamation mark. The *semantic relatedness* of a sense is calculated with respect to each sense of each term belonging to its own window and not to the whole document. The pseudo code is shown below in Algorithm 2 to determine the best sense of a term.

Algorithm 2 Beste Sense Algorithm

```
procedure Best_Sense(W_t)
    for each sense S_t,i of target word W_t do
        set Score_St_t,i = 0
        for each word W_j ≠ W_t in windows of context do
            init array temp_score
            for eachsenseS_j,k ofW_j do
                temp_score[j] = SRG(S_t,i, S_j,k)
            end for
            Score_S_t,i = Score_S_t,i + MAX(temp_score)
        end for
        if best_sense_score < Score_S_t,i then
            best_sense_score = Score_S_t,i
            best_sense = S_t,i
        end if
    end for
    return best_sense
end procedure
```

Fig. 3. Visual content of a document.
Source: Wikipedia.

The best sense of a term is that with the maximum score obtained by estimating the semantic relationship with all the other terms of a given context window.

3.3.2. Visual topic detection

The visual topic detection task exploits multimedia elements of a document, i.e. images, in order to determine its main topic. In our approach the visual topic detection is used to improve the performance of the whole framework. At this stage of our research, we are interested in the recognizing of multimedia descriptors to measure the similarity between an image in an analyzed document and images in our multimedia knowledge base. We use different descriptors based on local, global and deep features as described in Section 3.2. The performance of these descriptions have been tested following a strategy discussed in the next section.

Features extracted from the images in the document are matched with the ones contained in the multimedia knowledge base. The list of concepts is ranked according to the chosen similarity metric (e.g. the cosine similarity or the Manhattan distance). Each concept in the list has a score assigned to recognize the topic concept of the document. The score is computed according to the following formula:

$$TopicConcept = max((1 - distance) \cdot SenseCoverage \cdot depth \cdot freq) \quad (4)$$

where the *Sense Coverage* and the *depth* are the same as explained in 3.3.1; the frequency is used to consider the number of times a concept appears (due to the multiple visual representation of each concept); the distance is the value computed by the similarity metric used for the matching. It is normalized in the $[0, 1]$ range. If a concept appears more than one time we compute the average.

We provide also an example to better understand the Visual Topic Detection process above explained. At this stage, we consider media contents of documents (see Fig. 3).

A script, written in Python language, reads and processes the visual content to extract visual features from it. Then, a Cypher query is run on our multimedia knowledge graph to compute the similarity (i.e. cosine similarity) with features already stored in the knowledge base. Finally, we apply the formula given in Eq. (4), according to the same considerations discussed for the text topic detection. Consistency among textual and visual topic detection is always guaranteed by the fact that both processes are "semantically controlled" by the multimedia knowledge graph.

3.3.3. Combiner

The goal of the combination is to achieve better classification performances. In this paper, two classifiers has been proposed, one for the text-based topic detection and one for the visual-based topic detection. Several studies (Ghosh, Shankar, Bruzzone, & Meher, 2010; Kuncheva, 2001, 2004; Lam, 2000) show that different classifiers potentially offer complementary information models based on the patterns that must be classified and they could be used in a combined way to improve the overall system performance. The idea is not to use a single classifier but to consider several individual results in the final classification.

Table 6
Combination scores computation with average operator.

Concept	Score
hound.n.01	0.4
foxhound.n.01	0.2
beagle.n.01	**0.65**
dog.n.01	0.2
cat.n.01	0.05

The scores of each concept computed by the respective textual and visual topic detection are normalized with respect to their best score and they are normalized in the [0,1] interval. The combination of textual and visual classifications can be carried out according to various schemes, following Rinaldi (2013) we chose to use the *SUM* operator and the *OWA* operators (Yager & Kacprzyk, 2012). A brief explanation of such operators is provided in the following of this Section.

Considering the previous example of a document whose topic is "beagle", the combination technique can help in improving final results, taking out the best result from the two topic detection approaches. Text topic detection and visual topic detection outputs are a list of concepts each having a matching score. For the purpose of this example, we assume that the list of concepts for text topic detection is $[hound.n.01, beagle.n.01, foxhound.n.01]$ with values, normalized in the unit range, $[0.8, 0.7, 0.4]$, while visual topic detection output is $[beagle.n.01, dog.n.01, cat.n.01]$ with values $[0.6, 0.4, 0.3]$. By using an average operator as an example, we can combine these results together, obtaining the final result, as shown in Table 6.

The SUM function is one of the most adopted techniques for linear combinations of classifiers. There are different versions of it: *weighted sum*, *average* etc. and it provides better performance than other elementary functions due to its noise tolerance (Kittler, Hatef, Duin, & Matas, 1998).

Ordered Weighted averaging (OWA) operators provides a parameterized class of average aggregation operators. Some of the most notable, such as maximum, arithmetic mean, median and minimum, are members of this class. They have been widely used in computational intelligence due to their ability to model linguistically expressed aggregation instructions (Yager & Kacprzyk, 2012).

Formally an OWA operator of size n is a function $F : R_n \rightarrow R$ with a collection of associated weights $W = [w_1, \ldots, w_n]$ whose elements are in the unit range such that $\sum_{i=1}^{n} w_i = 1$. The function is defined as:

$$F(a_1, \ldots, a_2) = \sum_{j=1}^{n} w_j b_j \quad (5)$$

where b_j represents the jth value of the $\vec{a}$ vector ordered.

4. Test strategy and experimental results

In this section we show the experiments that have been carried out to test the following components of the proposed framework:

- Textual Topic Detection;
- Visual Topic Detection;
- Combined Topic Detection.

The system presented in this paper is highly generalizable due to the nature of the developed modules. It is possible, in fact, to submit to the system any multimedia document collection provided with a RDF description.

Table 7
DMOZ: general statistics.

Statistic item	Total
Languages	90
Documents	3 573 026
Categories	1 031 722
Categories levels	14
Editors	91 929

Table 8
DMOZ - categories statistics.

Level	Tot. categories
I	1
II	15
III	495
IV	4 384
V	19 042
VI	45 917
VII	48 101
VIII	62 970
IX	33 282
X	23 761
XI	21 997
XII	2 651
XIII	1 072
XIV	246
XV	40

Table 9
DMOZ - URLs/category.

DMOZ category	URLs/Category	URLs/Ground truth category
Arts	164 873	1312
Business	171 734	1208
Computers	78 994	1189
Games	28 260	1136
Health	41 905	1011
News	6391	1264
Science	79 733	1173
Shopping	60 891	1430
Society	169 054	1272
Sports	71 769	1125
	Tot. URLs 3 573 026	12 120

4.1. Dataset statistics

In this paper, we have used as data-set the multimedia content of *DMOZ* (DMOZ, 2002), one of the most popular and rich multilingual web directories with open content. The project's website makes it available as dumps in RDF format (Lassila, Swick, et al., 1998) from which it is possible, through a parser to extract the data of interest. The archive is made up of links to multimedia web content organized according to a hierarchy. Tables 7 and 8 show the statistics of the collection and of the categories contained in it, also organized in a hierarchical structure.

We choose DMOZ to have a real experimental scenario and to have a public and well know repository to compare our results with baselines.

The category at the first level represents the root of the hierarchy and it does not provide any informative contribute, hence it has been discarded. The *ground truth* has been built starting from a subset of documents belonging to categories at level two, shown in Table 9.

The list of URLs to download is submitted to the Scraper module with the parameter *text-only*; in this way only the textual part of each document will be retrieved if a valid response was received from the HTTP request. We use a part of the entire DMOZ repository due to different considerations, including the web-scraping policies and the presence of numerous dead links.

Table 10
General statistics test set.

Field	Value
Num doc.	1212
Avg length doc.	877,24
Max length doc.	2479
Min length doc.	341
Num. images	3643
Avg images	3

4.2. Test topic detection

The main aspect of our framework has been careful evaluated with the aim of showing the difference performances of the proposed methodology compared to the baselines in literature. Moreover, we highlight the differences between the combination of textual analysis and visual analysis and the single classifiers. In order to have a more robust evaluation of our system's performances, we also compare it with two reference algorithms used in the topic detection research field. These two algorithms are *LSA* and *LDA*. *Latent Semantic Analysis* (LSA) (Landauer et al., 1998), also known as *Latent Semantic Indexing* (LSI) is a technique of vectorial representation of a document though the *bag-of-words* model. *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003) is a text-mining model based on statistical models. More details have been provided in Section 2.

The annotation process shown in Table 11 also makes it easier to classify documents treated with LSA and LDA, because the two algorithms give a number of topics that represent the main topics of the analyzed corpus but which does not have relations with the DMOZ categories. On a total of 12 120 documents, 10 910 are used in order to create the topic modeling models used by LSA and LDA, while 1210 documents are used as test-sets for the entire system. A good test-set for the system must consist of documents that are complied with textual and visual analysis. In particular, an attempt was made to retrieve random documents from the web directory composed of a rich textual part and at least 3 images. The algorithm used by the DOM Parser module is listed below. It is used to recognize and store a "valid" multimedia document for the system purpose.

Algorithm 3 DOM Parser algorithm

```
1: procedure IsDocumentValid(htmlSource, numImages, numTokens)
2:     language = detectLanguage(htmlSource)
3:     if language == english then
4:         body = detectBody(htmlSource)
5:         for each tag t_i in body.tagList do
6:             if t_i in acceptedTags then
7:                 if t_i == imageTag then imageList += getTagContent(t_i)
8:                 else
9:                     textList += getWords(getTagContent(t_i))
10:                end if
11:            end if
12:        end for
13:        if len(imageList) >= numImages and len(textList) >= numTokens then
14:            return true, textList, imageList
15:        else
16:            return false
17:        end if
18:    else
19:        return false
20:    end if
21: end procedure
```

Table 11
Mapping of DMOZ categories and WordNet.

Cat.	Synset	Definition	Offset
Arts	art.n.01	The products of human creativity; works of art collectively	2743547
Business	commercial_enter-prise.n.02	The activity of providing goods and services involving financial and commercial and industrial aspects	1094725
Computers	computer.n.01	A machine for performing calculations automatically	3082979
Games	game.n.01	A contest with rules to determine a winner	455599
Health	health.n.01	A healthy state of well-being free from disease	14447908
News	news.n.01	Information about recent and important events	6642138
Science	science.n.01	A particular branch of scientific knowledge	5999797
Shopping	shopping.n.01	Searching for or buying goods or services	81836
Society	society.n.01	An extended social group having a distinctive cultural and economic organization	7966140
Sports	sport.n.01	An active diversion requiring physical exertion and competition	523513

Table 12
Detail of test for textual topic detection.

Algorithm	Accuracy	Num. correct
LSA	0,10	117
SEMREL	0,32	389
LDA	0,34	408
SEMREL WoC	0,41	502

Table 13
Detail of test for visual topic detection.

Feature	Accuracy	Num. correct
SIFT	0,29	471
PHOG	0,39	348
VGG16	0,70	846

We use English document for our experiments automatically handled with the help of a python library porting of the *Google* algorithm for language detection (Danilak, 2014, 2017). An additional library *BeautifulSoup* (Hajba, 2018) is used to perform the scraping of the page. The HTML tags of interest are selected and the related informative content is stored in the JSON format.

Table 10 shows the general statistics for the used test set.

The testing strategy proposed in this work involves the use of the multimedia knowledge graph for the topic classification. In order to use the same technique to have a reliable comparison for all the implemented algorithms, it is necessary to perform a manual mapping of the DMOZ categories used on the respective ImageNet synsets. In this way we create a ground truth by means of a pre-classified document directory and correspondence with a formal and well-known knowledge source. In particular, the following synsets have been chosen because they represent more rich informative content in the DMOZ categories with respect to generic and specialized topics.

Once documents have been retrieved from DMOZ, they are used as a corpus for generating LSA and LDA models. For this purpose, the models are created by applying the bag-of-words transformation to pre-processed documents by the textual pre-processing module and are created through a Python script with the help of the *gensim* library (Řehůřek & Sojka, 2010).

4.2.1. Textual topic detection

For the textual topic detection, the LSA and LDA models have been implemented and generated, as well as the proposed SEMREL algorithm in two variants. The first one consists in calculating the metric named *semantic relatedness grade* (SRG) (Rinaldi, 2009) of a sense related to a term semantically compared with all the terms of the whole document. The second one is between all terms belong to a window of context. In our case we have chosen as window of context a grammatical period. The end of a period is identified by punctuation symbols. Fig. 4 shows the obtained results accuracy and a summary is in Table 12.

The SEMREL algorithm, used with its window of context variant, provided the best results in term of accuracy for the textual topic detection, being followed by LDA and SEMREL, while LSA algorithm seems to not perform well as the other. A possible reason is that the results obtained are also dependent on the possibility of carrying out the mapping of the DMOZ categories on the concepts of the proposed ontology (11). For example, in case of LSA, the low accuracy value is also affected by the impossibility of associating some topics generated by the model with the corresponding WordNet synsets. We can argue that SEMREL have a better generalization of concept recognition taking out noise from specific datasets.

4.2.2. Visual topic detection

As explained in 3.3.2, given the list of concepts obtained from the matching, a semantic analysis is performed between them. If the document has more than one image, the algorithm selects the best candidate obtained from the highest assigned score computed by the similarity function. Three different visual topic detection configurations have been tested by means of the used features. In Fig. 5 the results obtained from the different strategies are shown. A summary is provided in Table 13.

The results show that the feature extracted from VGG16 has the highest accuracy, followed by PHOG and SIFT.

The results reflect our expectations, since PHOG and the feature extracted from VGG16 turn out to be the candidates to provide more accurate results in terms of feature matching. In particular, VGG16 features have a much higher discriminating power with respect to other descriptors. It should be noted that the price to pay for such accuracy is in the computational time. In fact the dimensionality of VGG16 features making computations slower when using such kind of feature. Hence, according to the application, we could prefer global or local features instead of deep features. Also, the higher accuracy of PHOG with respect to SIFT is not surprising, since PHOG is a global feature because it provides a single visual descriptor from an image, but it can be also seen as a local feature because it processes an image at different scales and resolutions. The final choice is given by the analysis of the combined strategy.

4.2.3. Combined topic detection

The aim of the combined topic detection is to assign a unique score depending on weights given to single textual and visual topic detection classifier.

The used combination techniques are the *SUM* and the *OWA* operators. Each proposed scheme has been tested with several weight combinations, except for the OWA schema with OWA operators which realizes a *fuzzy* logic approach. The whole test process has been performed on 48 combinations. The used schemes are shown in Table 14.

The *A* combination is related to the tests shown above, in which the visual topic detection has shown better performances. The *B* combination is an average of the computed scores while the *C* combination assigns more relevance to textual topic detection. A combination has also been tested with the OWA operators (the schema is indicated with the letter D) using as weights vector $\vec{w} = [0.65, 0.35]$ (see Fig. 6).

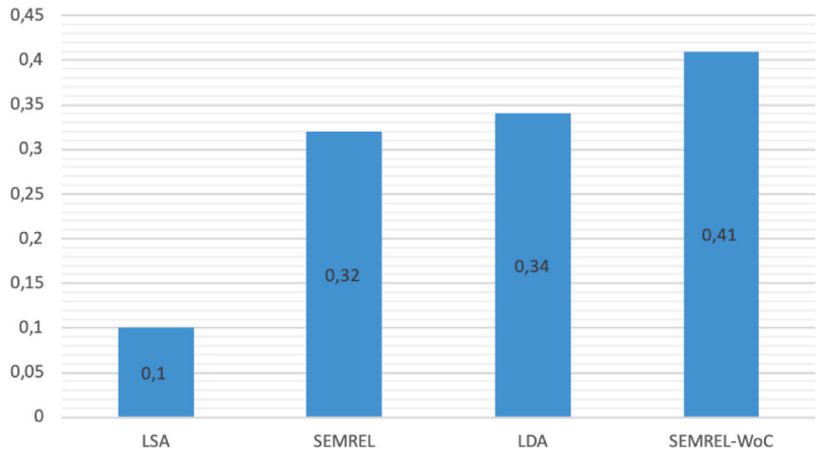

Fig. 4. Accuracy textual topic detection.

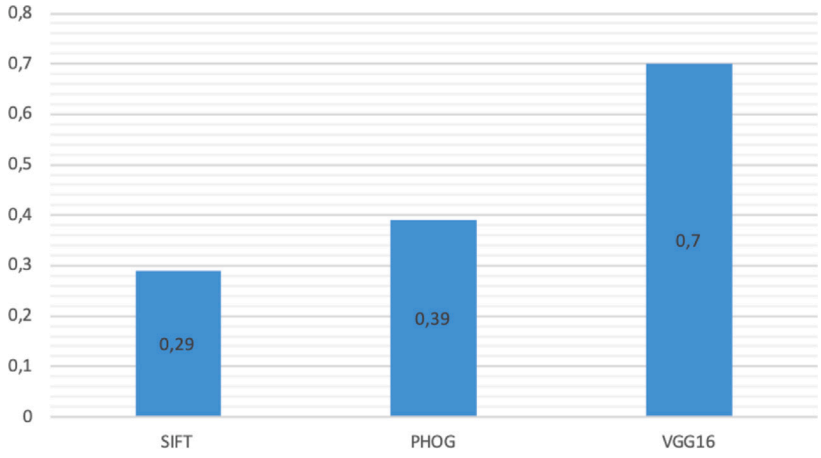

Fig. 5. Accuracy visual topic detection.

Table 14
Weights configurations for combined topic detection.

Combination	Weight textual TD	Weight visual TD
A	0.4	0.6
B	0.5	0.5
C	0.6	0.4
D	0.65	0.35

The figures show that the best combinations are the ones where the visual topic detection has a high weight compared to the textual topic detection (scheme A) and those that follow a fuzzy logic approach (schema D). The combinations based on deep features have a good accuracy and they give a significant contribution to the combination. The schemes with a high weight assigned to textual topic detection have, on the contrary, a low or similar accuracy with respect to single textual topic detection. In some cases the combination that assigned the same weight to textual and visual classifiers has low accuracy in the whole classification process.

Generally speaking, we argue that the contribution given by visual classifiers with a high accuracy is due to a better discrimination in the topic detection. We have these results because a specific concept (i.e. topic) is better represented by an image and, on the contrary, a specific term can have multiple meanings with a possible higher uncertainty in the topic detection task.

The best result is obtained by the combination between our proposed SEMREL algorithm using its version *Window of Context* and the visual topic detection based on the deep feature extracted from the VGG16 network. Even if the VGG16 descriptor needs a higher computational effort, we argue that the categorization process in our context of interest is an offline task and we are interested in a high accuracy. The Table 15 shows a complete overview of the results.

5. Conclusion and future works

In a world overwhelmed by data production, tools for document classification are needed. Such tools must be able to quickly provide efficient access to information during the retrieval process. In this article we proposed an approach to topic detection of multimedia documents. With respect to previous works, it combines both textual and visual features. Moreover it makes use of both statistics and semantics for multimedia web documents topic recognition. The use of semantic information, in particular, adds value to our results by leveraging them to an upper level, directly addressing a well-known issue, that is the *semantic gap*. Our work proposes several novelties mainly related to knowledge representation based on the property graph model to handle "big data" and techniques to analyze and model the topic detection process based on statistical models. The proposed approach gives high performance in different contexts because it can support the organization of a document collection and it can describe the main topic of a document by means of semantic multimedia analysis among concepts. Individual topic detection tasks have been evaluated through experiments and a discussion about their results have been provided. With these results, it has been possible to affirm that our proposed approach for text topic detection outperformed state-of-art algorithms, i.e. LSA, LDA, in this field. About the visual topic detection, different kind of

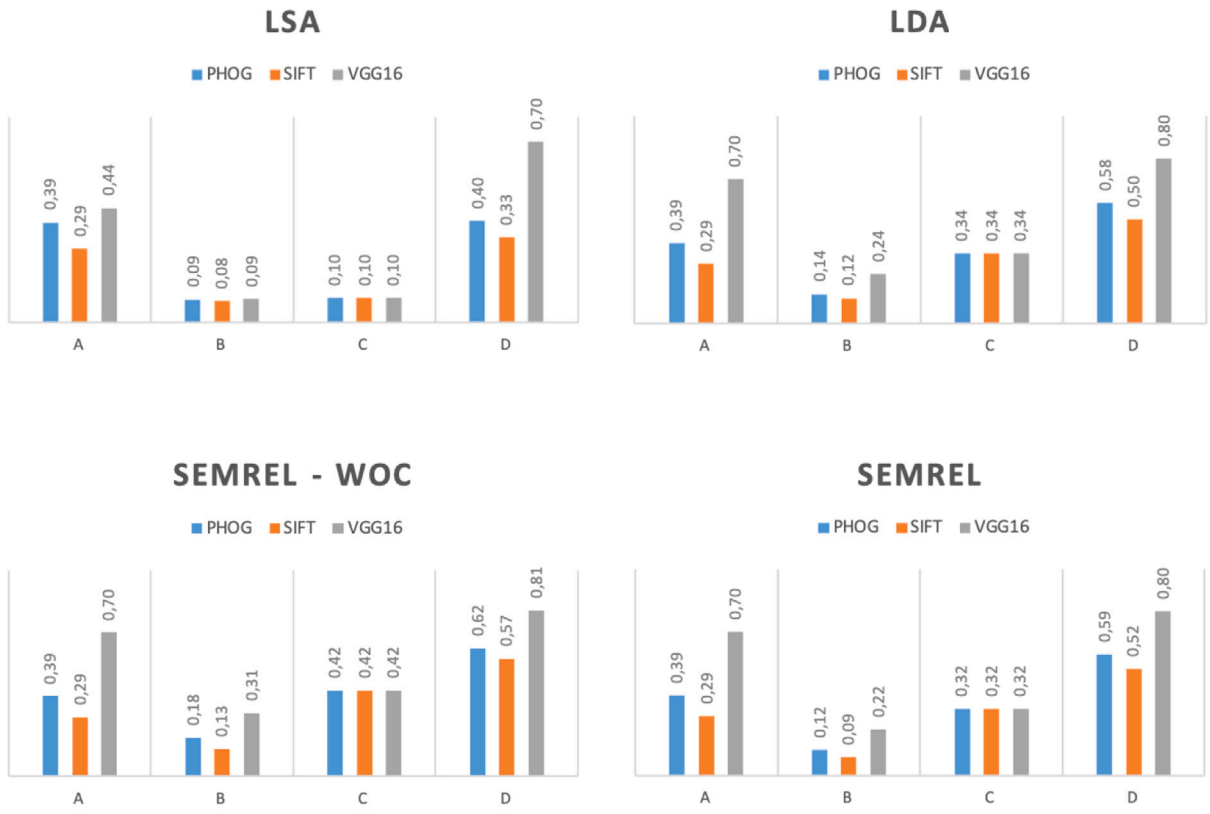

Fig. 6. Accuracy combined topic detection.

Table 15
Test details for combined topic detection.

Combination	Accuracy	Num. correct	Combination	Accuracy	Num. correct
SEMREL-WOC-VGG16-D	0.81	978	LDA-VGG16-D	0.80	966
SEMREL-VGG16-D	0.80	965	LSA-VGG16-D	0.70	852
LDA-VGG16-A	0.70	846	SEMREL-WOC-VGG16-A	0.70	846
SEMREL-VGG16-A	0.70	846	SEMREL-WOC-PHOG-D	0.62	750
SEMREL-PHOG-D	0.59	711	LDA-PHOG-D	0.58	708
SEMREL-WOC-SIFT-D	0.57	693	SEMREL-SIFT-D	0.52	629
LDA-SIFT-D	0.50	609	LSA-VGG16-A	0.44	539
SEMREL-WOC-PHOG-C	0.42	503	SEMREL-WOC-SIFT-C	0.42	503
SEMREL-WOC-VGG16-C	0.42	503	LSA-PHOG-D	0.40	480
LSA-PHOG-A	0.39	471	LDA-PHOG-A	0.39	471
SEMREL-WOC-PHOG-A	0.39	471	SEMREL-PHOG-A	0.39	471
LDA-PHOG-C	0.34	408	LDA-SIFT-C	0.34	408
LDA-VGG16-C	0.34	408	LSA-SIFT-D	0.33	402
SEMREL-PHOG-C	0.32	389	SEMREL-SIFT-C	0.32	389
SEMREL-VGG16-C	0.32	389	SEMREL-WOC-VGG16-B	0.31	371
LSA-SIFT-A	0.29	348	LDA-SIFT-A	0.29	348
SEMREL-WOC-SIFT-A	0.29	348	SEMREL-SIFT-A	0.29	348
LDA-VGG16-B	0.24	288	SEMREL-VGG16-B	0.22	270
SEMREL-WOC-PHOG-B	0.18	224	LDA-PHOG-B	0.14	171
SEMREL-WOC-SIFT-B	0.13	158	SEMREL-PHOG-B	0.12	149
LDA-SIFT-B	0.12	147	LSA-PHOG-C	0.10	117
LSA-SIFT-C	0.10	117	LSA-VGG16-C	0.10	117
LSA-VGG16-B	0.09	111	LSA-PHOG-B	0.09	108
SEMREL-SIFT-B	0.09	108	LSA-SIFT-B	0.08	100

descriptors have been tested. In particular, the features extracted from the activation layer of the VGG16 deep neural network model has shown best results. In addition, we have shown that combining the approach for textual topic detection with visual topic detection it is possible to improve the whole task, taking out the best results from both tasks. Since the system is modular and reusable, it is possible to extend the proposed architecture by introducing new modules or modifying the existing ones in order to implement different models with respect to the task of topic detection. The system has been fully tested against a general web document collection (i.e. *DMOZ*). Nonetheless, the design of the system allow the use of different collections of multimedia documents.

In the future research works we plan to extend the testing of our system with different document collections. In particular, given the difficulty in finding data-sets for web pages which contains both text and multimedia contents, one of the future work is to build such a data-set, also exploiting our knowledge base for topic identification. Moreover, new approaches to the topic detection will be investigated.

An interesting line of research consists in addressing the issue of multilingualism for improving textual topic detection performances. In particular, we will focus on the textual topic detection algorithm computational efficiency. We will investigate on novel methodologies and algorithms for visual topic detection to improve the state of art in this task.

CRediT authorship contribution statement

Antonio M. Rinaldi: Conceptualization, Methodology, Supervision. **Cristiano Russo:** Data curation, Writing - original draft, Writing - review & editing, Validation. **Cristian Tommasino:** Software, Visualization, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alghamdi, A. (2015). A survey of topic modeling in text mining. In *International journal of advanced computer science and applications*. IJACSA.

Alguliev, R. M., & Aliguliyev, R. M. (2005). Effective summarization method of text documents. In *Web intelligence, 2005. Proceedings. the 2005 IEEE/WIC/ACM international conference on* (pp. 264–271). IEEE.

Angles, R. (2018). The property graph database model. In *AMW*.

Asgari-Chenaghlu, M., Feizi-Derakhshi, M.-R., Balafar, M.-A., Motamed, C., et al. (2020). TopicBERT: A transformer transfer learning based memory-graph approach for multimodal streaming social media topic detection. arXiv preprint arXiv:2008.06877.

Bay, H., Ess, A., Tuytelaars, T., & Gool, L. V. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359, Similarity Matching in Computer Vision and Multimedia.

Bíró, I., Szabó, J., & Benczúr, A. A. (2008). Latent dirichlet allocation in web spam filtering. In *Proceedings of the 4th international workshop on adversarial information retrieval on the web* (pp. 29–32). ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bosch, A., Zisserman, A., & Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on image and video retrieval* (pp. 401–408). ACM.

Brooks, R. (1987). A hardware retargetable distributed layered architecture for mobile robot control. In *Proceedings. 1987 IEEE international conference on robotics and automation, Vol. 4* (pp. 106–110). IEEE.

Caldarola, E. G., & Rinaldi, A. M. (2016). Improving the visualization of wordnet large lexical database through semantic tag clouds. In *Big data (BigData congress), 2016 IEEE international congress on* (pp. 34–41). IEEE.

Cao, Y., Wang, C., Zhang, L., & Zhang, L. (2011). Edgel index for large-scale sketch-based image search. In *CVPR 2011*.

Castano, S., Ferrara, A., & Montanelli, S. (2003). H-match: an algorithm for dynamically matching ontologies in peer-based systems. In *Proceedings of the first international conference on semantic web and databases* (pp. 218–237). Citeseer.

Cavnar, W. B., Trenkle, J. M., et al. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, Vol. 161175*. Citeseer.

Chandiok, A., Prakash, A., Siddiqi, A., & Chaturvedi, D. (2020). Cognitive computing agent systems: An approach to building future real-world intelligent applications. *Computational Science and its Applications*, 257.

Danesi, M., & Perron, P. (1999). *Analyzing cultures: An introduction and handbook*. Indiana University Press.

Danilak, M. (2014). Langdetect: Language detection library ported from Google's language detection. See https://pypi.python.org/pypi/langdetect/ (accessed 19 January 2015).

Danilak, M. (2017). Langdetect 1.0. 7. *Python Package Index*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

DMOZ, D. (2002). Open directory project.

Gavrila, D. M., & Munder, S. (2007). Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, *73*(1), 41–59.

Ghosh, A., Shankar, B. U., Bruzzone, L., & Meher, S. K. (2010). Neuro-fuzzy-combiner: An effective multiple classifier system. *International Journal of Knowledge Engineering and Soft Data Paradigms*, *2*(2), 107–129.

Gopalakrishnan, K., Khaitan, S., Choudhary, A., & Agrawal, A. (2017). Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. *Construction and Building Materials*, *157*, 322–330.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220.

Hajba, G. L. (2018). *Website scraping with python: Using beautifulsoup and scrapy*. Apress.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 50–57).

Hu, X., & Wu, B. (2006). Automatic keyword extraction using linguistic features. In *Data mining workshops, 2006. ICDM workshops 2006. Sixth IEEE international conference on* (pp. 19–23). IEEE.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 216–223). Association for Computational Linguistics.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., et al. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211.

Khalid, H., & Wade, V. (2020). Topic detection from conversational dialogue corpus with parallel Dirichlet allocation model and elbow method. arXiv preprint arXiv:2006.03353.

Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239.

Kuncheva, L. I. (2001). Using measures of similarity and inclusion for multiple classifier fusion by decision templates. *Fuzzy Sets and Systems*, *122*(3), 401–407.

Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

Lam, A. (2000). Tacit knowledge, organizational learning and societal institutions: An integrated framework. *Organization Studies*, *21*(3), 487–513.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Lassila, O., Swick, R. R., et al. (1998). *Resource description framework (RDF) model and syntax specification*. Citeseer.

Li, W., Joo, J., Qi, H., & Zhu, S.-C. (2016). Joint image-text news topic detection and tracking by multimodal topic and-or graph. *IEEE Transactions on Multimedia*, *19*(2), 367–381.

Li, H., Ma, B., & Lee, C.-H. (2006). A vector space modeling approach to spoken language identification. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(1), 271–284.

Lin, Y.-z., Nie, Z.-h., & Ma, H.-w. (2017). Structural damage detection with automatic feature-extraction through deep learning. *Computer-Aided Civil and Infrastructure Engineering*, *32*(12), 1025–1046.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & E. Alsaadi, F. (2016). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*. IEEE.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, *13*(01), 157–169.

McGuinness, D. L., Van Harmelen, F., et al. (2004). OWL web ontology language overview. *W3C Recommendation*, *10*(10), 2004.

Melville, P., Gryc, W., & Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1275–1284). ACM.

Miller, G. A. (1995). WordNet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Miller, E. (1998). An introduction to the resource description framework. *Bulletin of the American Society for Information Science and Technology*, *25*(1), 15–19.

Mohanty, S. P., Hughes, D. P., & Salath, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, *7*, 1419.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, *61*(2), 217–235.

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, *8*(3), 489–508.

Prabhakar Kaila, D., Prasad, D. A., et al. (2020). Informational flow on Twitter–corona virus outbreak–topic modelling approach. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, *11*(3).

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). Valletta, Malta: ELRA, http://is.muni.cz/publication/884893/en.

Rinaldi, A. M. (2009). An ontology-driven approach for semantic information retrieval on the web. *ACM Transactions on Internet Technology (TOIT)*, *9*(3), 10.

Rinaldi, A. M. (2013). A multimodal content-based approach for web pages analysis. *International Journal of Knowledge Engineering and Data Mining*, *2*(4), 292–316.

Rinaldi, A. M. (2014). A multimedia ontology model based on linguistic properties and audio-visual features. *Information Sciences*, *277*, 234–246.

Rinaldi, A. M., & Russo, C. (2018a). A novel framework to represent documents using a semantically-grounded graph model. In *KDIR* (pp. 201–209).

Rinaldi, A. M., & Russo, C. (2018b). A semantic-based model to represent multimedia big data. In *Proceedings of the 10th international conference on management of digital ecosystems* (pp. 31–38). ACM.

Rinaldi, A. M., & Russo, C. (2018c). User-centered information retrieval using semantic multimedia big data. In *2018 IEEE international conference on big data (Big data)* (pp. 2304–2313). IEEE.

Russo, C., Madani, K., & Rinaldi, A. M. (2020). Knowledge acquisition and design using semantics and perception: A case study for autonomous robots. *Neural Processing Letters*, 1–16.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Santini, M., & Rosso, M. (2008). Testing a genre-enabled application: A preliminary assessment. In *FDIA'08, Proceedings of the 2nd BCS IRSG conference on future directions in information access* (p. 7). Swindon, UK: BCS Learning & Development Ltd..

Sarica, S., Luo, J., & Wood, K. L. (2020). TechNet: Technology semantic network based on patent data. *Expert Systems with Applications*, *142*, Article 112995.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sowa, J. F. (1987). *Semantic networks*. Citeseer.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*(3), 293–300.

Takçı, H., & Güngör, T. (2012). A high performance centroid-based classification approach for language identification. *Pattern Recognition Letters*, *33*(16), 2077–2084.

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, *34*(4), 2622–2629.

Wang, J., & Hua, X.-S. (2011). Interactive image search by color map. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(1), 1–23.

Webber, J. (2012). A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on systems, programming, and applications: Software for humanity* (pp. 217–218).

Wei, Y. (2012). An iterative approach to keywords extraction. In *International conference in swarm intelligence* (pp. 93–99). Springer.

Woods, W. A. (1988). What's in a link: Foundations for semantic networks. *Readings in Cognitive Science*, 102–125.

Xu, S., Yang, S., & Lau, F. C.-M. (2010). Keyword extraction and headline generation using novel word features. In *AAAI* (pp. 1461–1466).

Yager, R. R., & Kacprzyk, J. (2012). *The ordered weighted averaging operators: theory and applications*. Springer Science & Business Media.

Zhang, H. (2004). The optimality of naive Bayes. *AA*, *1*(2), 3.

Zhi, T., Duan, L., Wang, Y., & Huang, T. (2016). Two-stage pooling of deep convolutional features for image retrieval. In *2016 IEEE international conference on image processing (ICIP)* (pp. 2465–2469).