Web Document Categorization Using Knowledge Graph and Semantic Textual Topic Detection

Antonio M. Rinaldi(✉), Cristiano Russo, and Cristian Tommasino

Department of Electrical Engineering and Information Technologies,
University of Napoli Federico II, 80125 Via Claudio, 21, Napoli, Italy
{antoniomaria.rinaldi,cristiano.russo,cristian.tommasino}@unina.it

Abstract. In several contexts, the amount of available digital documents increases every day. One of these challenging contexts is the Web. The management of this large amount of information needs more efficient and effective methods and techniques for analyzing data and generate information. Specific application as information retrieval systems have more and more high performances in the document seeking process, but often they lack of semantic understanding about documents topics. In this context, another issue arising from a massive amount of data is the problem of information overload, which affects the quality and performances of information retrieval systems. This work aims to show an approach for document classification based on semantic, which allows a topic detection of analyzed documents using an ontology-based model implemented as a semantic knowledge base using a No SQL graph DB. Finally, we present and discuss experimental results in order to show the effectiveness of our approach.

1 Introduction

The widespread diffusion of new communication technologies, such as the Internet together with the development of intelligent artificial systems capable of producing and sharing different kinds of data, have led to a dramatic increasing of the number of available information. One of the main goal in this context is to transform heterogeneous and unstructured data into useful and meaningful information through the use of Big Data, deep neural networks and the myriad of applications that derive from their implementations. For this purpose, documents categorization and classification is an essential task in the information retrieval domain, strongly affecting user perception [16]. The goal of classification is to associate one or more classes to a document, easing the management of a document collection. The techniques used to classify a document have been widely applied to different contexts paying attention to the semantic relationships especially between terms and the represented concepts [27,28]. The use of semantics in the document categorization task has allowed a more accurate detection of topics concerning classical approaches based on raw text and meaningless label [24]. Techniques relying on semantic analysis are often based on the idea of

O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12951, pp. 40–51, 2021.
https://doi.org/10.1007/978-3-030-86970-0_4

semantic network (SN) [32]. Woods [37] highlighted the lack of a rigorous definition for semantic networks and their conceptual role. In the frame of this work, we will refer to a semantic network as a graph entity which contains information about semantic and/or linguistic relationships holding between several concepts. Lately, semantic networks have been often associated to ontologies which are now a keystone in the field of knowledge representation, integration and acquisition [26,29–31]. Moreover, ontologies are designed to be machine-readable and machine-processable. Over the years, the scientific community provided many definitions of ontologies. One of the most accepted is in [11]. It is possible to represent ontologies into graphs and vice versa, with this duality making them interchangeable. The use of graphs and analysis metrics permits us to have a fast retrieval of information and for finding new patterns of knowledge hard to recognize. Topic detection and categorization are crucial task which allows quick access to contents in a document collections when used in an automatic way. A disadvantage of many classification methods is that they treat the categorization structure without considering the relationships between categories. A much better approach is to consider that structures, either hierarchical or taxonomic, constitute the most natural way in which concepts, subjects or categories are organized in practice [1].

The novelty of the proposed work has to be found in the way we combine statistical information and natural language processing. In particular, the approache uses an algorithm for word sense disambiguation based on semantic analysis, ontologies and semantic similarity metrics. The core is a knowledge graph which represents our semantic networks (i.e. ontology). It is used as a primary source for extracting further information. It is implemented by means of a NoSQL technology to perform a "semantic topic detection".

The paper is organized as follows: in Sect. 2 we provide a review of the literature related to Topic Modeling and Topic Detection techniques and technologies; Sect. 3 introduces the approach along with the general architecture of the system and the proposed textual classification methodology; in Sect. 4 we present and discuss the experimental strategy and results; lastly, Sect. 5 is devoted to conclusions and future research.

2 Related Works

This section analyzes relevant recent works related to textual topic detection, as well as the differences between our approach and the described ones. Over the years, the scientific community proposed several methodologies, hare grouped according to the main technique used. The goal of approaches based on statistics is to identify the relevance of a term based on some statistical properties, such as TF-IDF [33], N-Grams [8], etc. *Topic modeling* [20] instead is an innovative and widespread analytical method for the extraction of co-occurring lexical clusters in a documentary collection. In particular, it makes use of an ensemble of unsupervised text mining techniques, where the approaches are based on probabilities. The authors in [21] described a probabilistic approach for Web page classification, they propose a dynamic and hierarchical classification system that is capable of adding new categories, organizing the Web pages into a tree structure and classifying them by searching through only one path of the tree structure. Other approaches use *features* based on linguistic, syntactic, semantic, lexical properties. Hence, they are named linguistic approaches. Similarity functions are employed to extract representative keywords. Different machine learning techniques, such as Support Vector Machine [35], Naive Bayes [39] and others are used. The keyword extraction is the result of a trained model able to predict significant keywords. Other approaches attempt to combine the above-cited ones in several ways. Other parameters such as *word position*, *layout feature*, HTML tags, etc. are also used. In [13], the authors use an approach based on machine learning techniques in combination with semantic information, while in [18] co-occurrence is employed for the derivation of keywords from a single document. In [12], the authors use linguistic features to represent term relevance considering the position of a term in the document and other researches [25] build models of semantic graphs for representing documents. In [36], the authors presented an iterative approach for keywords extraction considering relations at different document levels (words, sentences, topics). With such an approach a graph containing relationships between different nodes is created, then the score of each keyword is computed through an iterative algorithm. In [2], the authors analyzed probabilistic models for topic extraction. Xu et al. [38] centered their research on topic detection and tracking but focusing on online news texts. The authors propose a method for the evolution of news topics over time in order to track topics in the news text set. First, topics are extracted with LDA (latent Dirichlet allocation) model from news texts and the Gibbs Sampling method is used to define parameters. In [34] an extended LDA topic model based on the occurrence of topic dependencies is used for spam detection in short text segments of web forums online discussions. Khalid et al. [14] use parallel dirichlet allocation model and elbow method for topic detection from conversational dialogue corpus. Bodrunova et al. [5] propose an approach based on sentence embeddings and agglomerative clustering by Ward's method. The Markov stopping moment is used for optimal clustering. Prabowo et al. [22] describe a strategy to enhance a system called ACE (Automatic Classification Engine) using ontologies. The authors focus on the use of ontologies for classifying Web pages concerning the Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) schemes using weighted terms in the Web pages and the structure of domain ontologies. The association between significant conceptual instances into their associated class representative(s) is performed using an ontology classification scheme mapping and a feed-forward network model. The use of ontologies is also explored in [17]. The authors propose a method for topic detection and tracking based on an event ontology that provides event classes hierarchy based on domain common sense.

In this paper, we propose a semantic approach for document classification. The main differences between our approach and the other presented so far are in the proposing of a novel algorithm for topic detection based on semantic information extracted from a general knowledge base for representing the user domains of interest and the fully automatization of our process without a learning step.

3 The Proposed Approach

In this section, we provide a detailed description of our approach for topic detection. The main feature of our methodology is its ability to combine both statistical information, natural language processing and several technologies to categorize documents using a comprehensive semantic analysis, which involves ontologies and metrics based on semantic similarity. To implement our approach, we follow a modular framework for document analysis and categorization. The framework makes use of a general knowledge base, where textual representation of semantic concepts are stored.

3.1 The Knowledge Base

We realized a general knowledge base using an ontology model proposed and implemented in [6,7]. The database is realized by means of a NoSQL graph technology. From an abstract, conceptual point of view, the model representation is based on *signs*, defined in [9] as "*something that stands for something, for someone in some way*". These signs are used to represent concepts. The model structure is composed of a triple $<S, P, C>$ 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. We propose an approach focused on the use of textual representations and based on the semantic dictionary WordNet [19]. According to the terminology used in the ontology model, the textual representations are our signs. The ontology is defined using the DL version of the *Web Ontology Language*(OWL), a markup language that offers a high level of expressiveness preserving completeness and computational decidability. The model can be seen as a top-level ontology, since it contains a very abstract definition for its classes. The model and the related knowledge graph have been implemented in Neo4J *graph-db* using the *property-graph-model* [3].

Figure 1 shows a part of our knowledge graph to put in evidence the complexity of the implemented graph for a sake of clarity. It is composed of near 15,000 nodes and 30,000 relations extracted from our knowledge base.

3.2 The Topic Detection Strategy

Our novel strategy for textual topic detection is based on an algorithm called SEMREL. Its representation model is the classical *bag-of-words*. Once a document is cleaned, i.e. unnecessary parts are removed, the *tokenization* step allows to obtain a list of terms in the document. Such a list of terms is the input for a *Word Sense Disambiguation* step that pre-processes the list assigning the right meaning to each term. Then, Semantic Networks dynamically extracted

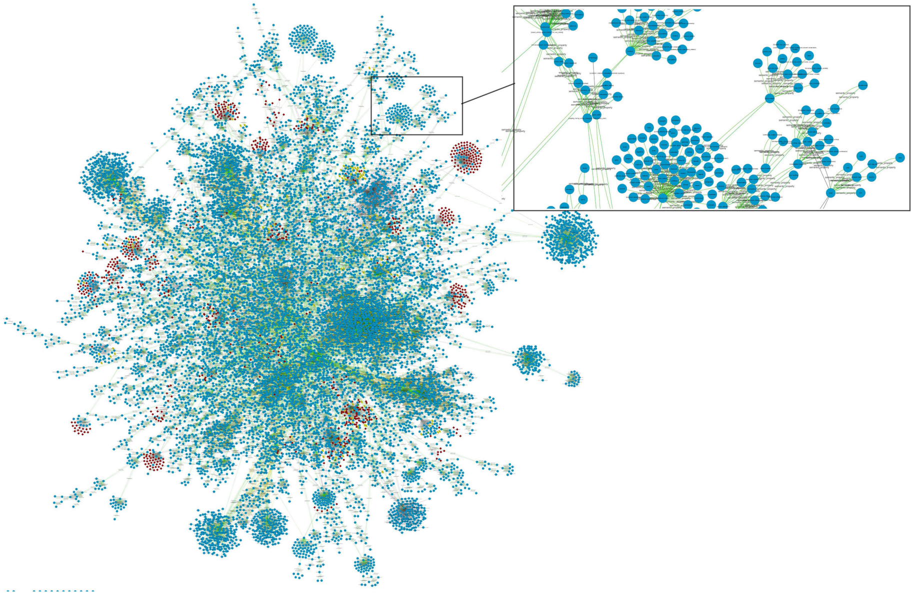

Fig. 1. Knowledge graph excerpt with 30000 edges and about 15000 nodes

from our knowledge base are generated for all the terms, in the *SN Extractor* step. Common nodes between an SN resulting from each concept and SNs from other concepts are used to compute their intersections. The common nodes correspond to the degree of representation of the concept considered with respect to the entire document. This measure is indicated as *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. It is computed as the number of hops to the root of our knowledge base considering only the hypernymy relationships. The *TopicConcept* is the one with the best trade-off between the *SenseCoverage* and the *Depth*. The formula used for calculating the topic concept of a given document is shown in Eq. 1.

$$TopicConcept = max(depth(C_i) * Coverage(C_i)) \tag{1}$$

where C_i is the *i-th* concept resulting from the WSD step. Only concepts in the *noun* lexical category are considered from the WSD list, because in the authors' opinion they are more representative to express the topic of a document.

In the Algorithm 1, we show the logic used to find the topic concept.

The WSD attempts to palliate the issue related to term *polysemy*. Indeed, it tries to "sense" the correct meaning of a term 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 computes their *semantic relatedness* [23].

Algorithm 1. Topic Concept Algorithm

```
procedure TopicConcept($ConceptList$)
    $BestConceptScore = 0$
    for each concept $C_i$ in ConceptList (after WSD) do
        $ScoreC_i = 0$
        $SN\_C_i = BuildSN(C_i)$
        $CoverC_i = 0$
        for each concept $C_j \neq C_i$ in ConceptList do
            $SN\_C_j = BuildSN(C_j)$
            $NumberOfCommonConcept = Match(SN\_C_i, SN\_C_j)$
            $CoverC_i = CoverC_i + NumberOfCommonConcept$
        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
```

This metric is based on a combination of the best path between pairs of terms and the depth of their Lowest Common Subsumer, expressed as the number of hops to the root of our knowledge base using hypernymy relationships.

The best path is calculated as follows:

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

where l is the best path length between the terms w_i and w_j, $h_j(w_i, w_j)$ corresponds to the number of hops of the *j-th* path and σ_i corresponds to the weight of the *i-th* edge of the *j-th* path. The weights σ_i are assigned to the properties of the ontological model described in Sect. 3.1 to discriminate the expressive power of relationships and they are set by experiments.

The depth factor is used to give more importance to specific concepts (low level and therefore with high depth) than generic ones (low depth). A non-linear function is used to scale the contribution of the sub-ordinates concepts in the upper level and increase those of a lower ones. The metric is normalized in the range $[0, 1]$ (1 when the length of the path is 0 and 0 when the length go to infinite).

The Semantic Relatedness Grade of a document is then calculated as:

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

where (w_i, w_j) are pairs of terms in v, $d(w_i, w_j)$ is the number of hops from the w_i, w_j subsumer to the root of the WordNet hierarchy considering the IS-A relation, α and β are parameters whose values are set by experiments.

The WSD process calculates the score for each sense of the considered term 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 terms in the document.

The best sense recognition is shown in the Algorithm 2.

Algorithm 2. Best 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 $W_j \neq W_t$ 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
```

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

3.3 The Implemented System

The system architecture is shown in Fig. 2. It is composed by multiple modules which are responsible of managing several tasks.

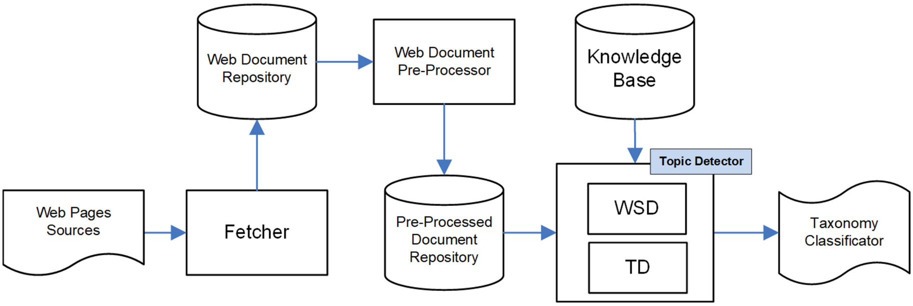

Fig. 2. The system architecture

The Web Documents can be fetched from different data sources by means of the Fetcher module and stored in the Web document repository. The textual information is first pre-processed. Cleaning operations are carried out by the Document Pre-Processor module. Such operations are: (i) tags removing, (ii) stop words deleting, (iii) elimination of special characters, (iv) stemming. The Topic Detection module uses an algorithm based on text analysis to address the correct topic of a document and our graph knowledge base. It is based on WSD and TD tasks based on the algorithm previously discussed. It is able to classify a document by the recognition of its main topic. Topic Detection result is the input of the Taxonomy Classificator used to create, with the help of our knowledge base, a hierarchy beginning from a concept. The proposed metric and approach have been compared with baselines and the results are shown in the next section.

4 Test Strategy and Experimental Results

In order to measure the performances of our framework we have carried out several experiments, which are discussed in the following. First we compare it with two reference algorithms widely used in the topic detection research field in order to have a more robust and significant evaluation: *LSA* [15] and *LDA* [4]. *Latent Semantic Analysis* (LSA), also known as *Latent Semantic Indexing* (LSI), is based on a vectorial representation of a document though the *bag-of-words* model. *Latent Dirichlet Allocation* (LDA) is a text-mining technique based on statistical models.

One of the remarkable feature of this system is that it is highly generalizable thanks to the development of autonomous modules. In this paper, we have used the textual content of *DMOZ* [10], one of the most popular and rich multilingual web directories with open content. The archive is made up of links to web content organized according to a hierarchy. The reason why we choose DMOZ lays into the fact that we want to compare our results with baselines. This way we can test against a real experimental scenario by using a public and well know repository. The category at the top level is the root of the DMOZ hierarchy. Since this is not informative at all, it has been discarded. Then we built a *ground truth* has been built considering a subset of documents from categories placed at the second level. These are shown in Table 1 together with statistics for the used test set.

The list of URLs is submitted to our fetcher to download the textual content. The restriction to a subset of DMOZ was necessary, due to the presence of numerous dead links and textual information. On a total of 12120 documents, we selected 10910 of them to create the topic modeling models used by LSA and LDA, while 1210 documents are used as test-set. The testing procedure employed in this paper uses our knowledge graph for the topic classification task. In order to have a fair and reliable comparison with all implemented algorithms, the same technique must be used, hence we need to perform a manual mapping of the used DMOZ categories to their respective WordNet synonyms. In this way, we create a ground truth using a pre-classified document directory (i.e., DMOZ) through a mapping with a formal and well-known knowledge

Table 1. DMOZ - URLs/category

DMOZ category	URLs/Category	URLs/Ground truth category
Arts	164 873	1 312
Business	171 734	1 208
Computers	78 994	1 189
Games	28 260	1 136
Health	41 905	1 011
News	6 391	1 264
Science	79 733	1 173
Shopping	60 891	1 430
Society	169 054	1 272
Sports	71 769	1 125
	Tot. URLs 3 573 026	12 120

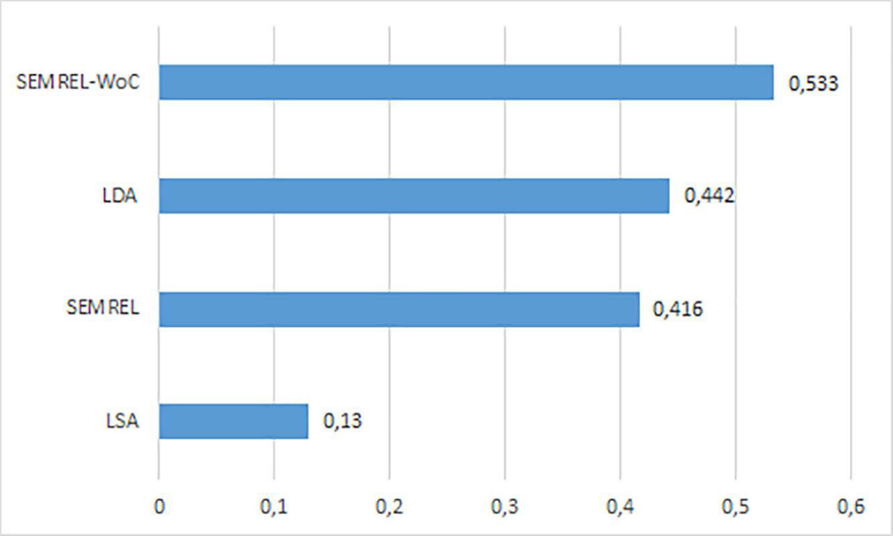

Fig. 3. Accuracy textual topic detection

source (i.e., WordNet). The annotation process also facilitates the classification of documents by other algorithms, e.g. LSA and LDA, because they give several topics that represent the main topics of the analyzed collection without dealings with the DMOZ categories. The central facet of our framework has been carefully evaluated to show the distinguishable performances of the proposed methodology. 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 computing the (SRG) of a sense related to a term semantically compared with all the terms of the whole document. The second one is performed dividing a document in grammatical periods, defined by the punctuation marks dot, question mark and exclamation mark (i.e. wondows of context). The *semantic relatedness* of a concept is calculated considering each sense of a term belonging to its window of context. Figure 3 shows the obtained results accuracy.

We argue that these results depend to the impossibility of mapping some topics generated by LSA or LDA with the corresponding WordNet sysnset. This issue dosn't allow an accurate topic detection due to the dependency of these models to the data set. On the other hand, SEMREL have a better concept recognition taking out noise from specific datasets.

5 Conclusion and Future Works

In this paper, we have proposed a semantic approach based on a knowledge graph for web document textual topic detection. For this purpose, a word sense disambiguation algorithm has been implemented, and semantic similarity metrics have been used. The system has been fully tested with a standard web document collection (i.e. *DMOZ*). The design of the system allows the use of different collections of documents. The evaluation of our approach shows promising results, also in comparison with state-of-art algorithms for textual topic detection. Our method has some limitations due to the lack of knowledge in several conceptual domains in our knowledge base (i.e. WordNet). In future works, we are interested in the definition of automatic techniques to extend our knowledge base with additional multimedia information and domain specific ontologies. Moreover, we want investigate on the novel methodologies to improve the performance of the topic detection process exploiting multimedia data considering new metrics to compute semantic similarity. Other aspects to point out are the computational efficiency of our approach and additional testing with different document collections.

References

1. Albanese, M., Picariello, A., Rinaldi, A.: A semantic search engine for web information retrieval: an approach based on dynamic semantic networks. In: Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2004)
2. Alghamdi, A.: A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl. IJACSA (2015)
3. Angles, R.: The property graph database model. In: AMW (2018)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
5. Bodrunova, S.S., Orekhov, A.V., Blekanov, I.S., Lyudkevich, N.S., Tarasov, N.A.: Topic detection based on sentence embeddings and agglomerative clustering with Markov moment. Future Internet **12**(9), 144 (2020)

6. Caldarola, E.G., Picariello, A., Rinaldi, A.M.: Experiences in wordnet visualization with labeled graph databases. In: Fred, A., Dietz, J.L.G., Aveiro, D., Liu, K., Filipe, J. (eds.) IC3K 2015. CCIS, vol. 631, pp. 80–99. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-52758-1_6
7. Caldarola, E.G., Picariello, A., Rinaldi, A.M.: Big graph-based data visualization experiences: the wordnet case study. In: 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), vol. 1, pp. 104–115. IEEE (2015)
8. Cavnar, W.B., Trenkle, J.M., et al.: N-gram-based text categorization. In: Proceedings of SDAIR 1994, 3rd Annual Symposium on Document Analysis and Information Retrieval, vol. 161175. Citeseer (1994)
9. Danesi, M., Perron, P.: Analyzing Cultures: An Introduction and Handbook. Indiana University Press, Bloomington (1999)
10. DMOZ: Dmoz website. http://dmoz-odp.org/
11. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum. Comput. Stud. **43**(5–6), 907–928 (1995)
12. Hu, X., Wu, B.: Automatic keyword extraction using linguistic features. In: Sixth IEEE International Conference on Data Mining Workshops, ICDM Workshops 2006, pp. 19–23. IEEE (2006)
13. Hulth, A.: 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 (2003)
14. Khalid, H., Wade, V.: Topic detection from conversational dialogue corpus with parallel dirichlet allocation model and elbow method. arXiv preprint arXiv:2006.03353 (2020)
15. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. **25**(2–3), 259–284 (1998)
16. Liaw, S.S., Huang, H.M.: An investigation of user attitudes toward search engines as an information retrieval tool. Comput. Hum. Behav. **19**(6), 751–765 (2003)
17. Liu, W., Jiang, L., Wu, Y., Tang, T., Li, W.: Topic detection and tracking based on event ontology. IEEE Access **8**, 98044–98056 (2020)
18. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. Int. J. Artif. Intell. Tools **13**(01), 157–169 (2004)
19. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
20. Papadimitriou, C.H., Raghavan, P., Tamaki, H., Vempala, S.: Latent semantic indexing: a probabilistic analysis. J. Comput. Syst. Sci. **61**(2), 217–235 (2000)
21. Peng, X., Choi, B.: Automatic web page classification in a dynamic and hierarchical way. In: Proceedings of 2002 IEEE International Conference on Data Mining, pp. 386–393. IEEE (2002)
22. Prabowo, R., Jackson, M., Burden, P., Knoell, H.D.: Ontology-based automatic classification for web pages: design, implementation and evaluation. In: Proceedings of the Third International Conference on Web Information Systems Engineering, WISE 2002, pp. 182–191. IEEE (2002)
23. Rinaldi, A.M.: An ontology-driven approach for semantic information retrieval on the web. ACM Trans. Internet Technol. (TOIT) **9**(3), 10 (2009)
24. Rinaldi, A.M.: Using multimedia ontologies for automatic image annotation and classification. In: 2014 IEEE International Congress on Big Data, pp. 242–249. IEEE (2014)

25. Rinaldi, A.M., Russo, C.: A novel framework to represent documents using a semantically-grounded graph model. In: KDIR, pp. 201–209 (2018)
26. Rinaldi, A.M., Russo, C.: 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 (2018)
27. Rinaldi, A.M., Russo, C.: User-centered information retrieval using semantic multimedia big data. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 2304–2313. IEEE (2018)
28. Rinaldi, A.M., Russo, C.: Using a multimedia semantic graph for web document visualization and summarization. Multimedia Tools Appl. **80**(3), 3885–3925 (2021)
29. Rinaldi, A.M., Russo, C., Madani, K.: A semantic matching strategy for very large knowledge bases integration. Int. J. Inf. Technol. Web Eng. (IJITWE) **15**(2), 1–29 (2020)
30. Russo, C., Madani, K., Rinaldi, A.M.: Knowledge acquisition and design using semantics and perception: a case study for autonomous robots. Neural Process. Lett. 1–16 (2020)
31. Russo, C., Madani, K., Rinaldi, A.M.: An unsupervised approach for knowledge construction applied to personal robots. IEEE Trans. Cogn. Dev. Syst. **13**(1), 6–15 (2020)
32. Sowa, J.F.: Principles of Semantic Networks: Explorations in the Representation of Knowledge. Morgan Kaufmann, Burlington (2014)
33. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **28**(1), 11–21 (1972)
34. Sun, Y.: Topic modeling and spam detection for short text segments in web forums. Ph.D. thesis, Case Western Reserve University (2020)
35. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
36. Wei, Y.: An iterative approach to keywords extraction. In: Tan, Y., Shi, Y., Ji, Z. (eds.) ICSI 2012. LNCS, vol. 7332, pp. 93–99. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31020-1_12
37. Woods, W.A.: What's in a link: Foundations for semantic networks. Read. Cogn. Sci. 102–125 (1988)
38. Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C., Yao, H.: Research on topic detection and tracking for online news texts. IEEE Access **7**, 58407–58418 (2019)
39. Zhang, H.: The optimality of Naive Bayes. AA **1**(2), 3 (2004)