A Document Visualization Strategy Based on Semantic Multimedia Big Data

Antonio M. Rinaldi[1,2](✉)

[1] Department of Electrical Engineering and Information Technologies,
University of Naples Federico II, Naples, Italy
antoniomaria.rinaldi@unina.it

[2] IKNOS-LAB Intelligent and Knowledge Systems,
University of Naples Federico II, LUPT, Naples, Italy

Abstract. The integration of semantic web and big data is a key factor in the definition of efficient model to represent knowledge and implement real world applications. In this paper we present a multimedia knowledge base implemented as a semantic multimedia big data storing semantic and linguistic relations between concepts and their multimedia representations. Moreover, we propose a document visualization strategy based on statistical and semantic analysis of textual and visual contents. The proposed approach has been implemented in a tool, called Semantic Tag Cloud, whose task is to show in a concise way the main topic of a document using both textual features and images. We also propose a case study of our approach and an evaluation from a user perception point of view.

Keywords: Knowledge base · Semantic big data · Ontology · WordNet · ImageNet · graphDB · Semantic tag cloud

1 Introduction

In the big data years, the use of approaches to capture, store and analyze data is a crucial factor in the implementation of knowledge based systems and smart applications. In this context, user oriented system use formal knowledge representations based on different symbols to easily interact with humans and, if it is represented using well know formalism like ontologies [26], it can be used by machines. On the other hand, novel technologies as Big Data give useful tools to manage high volume of data together with other dimensions related to changing data velocity and variety [16]. In the last years, a novel research area was arising from the Big Data paradigm and the Semantic Web vision called Semantic Big Data [35]. The union of these approaches give use novel strategies for managing and analyzing the large amount of data and useful tools to transform data in information and knowledge.

Our approach is based on the retrieval of knowledge represented in different multimedia forms [3] linked with semantic and linguistic relationships. The semantic information is extracted from a knowledge base in which concepts are

C. Esposito et al. (Eds.): I-SPAN 2019, CCIS 1080, pp. 43–57, 2019.
https://doi.org/10.1007/978-3-030-30143-9_4

represented in various multimedia forms and it is implemented by means of a big data. Our semantic multimedia big data (SMBD) is used to implement a the implementation of a visual smart interface.

Generally speaking, it is a visual semantic tag cloud where concepts have both a text and visual representations. The textual view uses a set of terms in which their characteristics are used (e.g. size, color and font) to indicate the important relations between terms and documents [44]. On the other hand, the visual representation shows the multimedia form (i.e. images) related to a concept. Tag clouds derive from the collaborative tagging paradigm [27] used in many social networks, forums and users' communities. In [6] the authors argue that if terms for tagging are recognized only in according to their frequency, several issues related to high semantic density because few topics will tend to dominate the entire visualization and less important terms will vanish. Tag clouds synthesize document collections by providing a brief summary of their information content. The purpose of summarization is to reduce the length of a document by creating a summary that preserves the most important points in the original document. In this way a user can preliminary understand the main topic of retrieved documents and decide if it is relevant or not for his/her research purpose [45,46].

In our paper, a novel method based on the combination of a textual and visual analysis of the original document to better understand the document content is presented.

The paper is organized as follows: in Sect. 2 we present an analysis of works related to semantic big data and document visualization; Sect. 3 is focused on the SMBD model and it implementation and the proposed strategy for document visualization is described in Sects. 4 and 5 shows a use case example of our strategy; eventually, Sect. 6 is devoted to the presentation of conclusion and future works.

2 Related Work

The evolution of Big Data research in the last years has followed different lines. In this section, we highlight some novel directions for the integration between Big Data and semantic web technologies. Moreover, some works related to document visualization using tag clouds also are presented.

A first formal definition of Big Data is given in [23]. In this paper the authors define Big Data as an information resource with specific dimensions related to a high *Volume*, *Velocity* and *Variety* of data. This asset requires specific techniques, methodologies and technologies to analyze data and transform it into *Value*. The Big Data dimensions can be extended with another one if we consider the *Veracity* [7]. Several surveys have been proposed in literature tho give a comprehensive classification of big data components both from a theoretic and technological point of view [17,34]. *Volume* and *Velocity* dimensions have been extensive studied during the last years but a methodological attention to *Variety* is arose only lately. In [35] a process called Big Data "Semantification" is presented and it is based on a methodology to manage heterogeneous Big Data using semantic technologies. In this way, a non-semantic Big Data is annotated by means of RDF vocabularies. The need of merging Big Data technologies in legacy systems using semantic-based approaches is very important issue and it emerges in different contexts [25]. In [1] a chain of NLP modules within virtual machines integrated in a distributed architecture for scaling up text analysis is proposed. Great efforts to enable social applications are based on the use of semantic and ontologies approaches to implement social multimedia big data [7,18]. The authors describe an online news management system [52] implemented following a link network model. An extract-transformation-load system to integrate different data sets by means of ad hoc semantic model is presented in [5]. It produces semantic linked data complied with the data model. A framework to query big data sources using Resource Description Framework (RDF) representation is presented in [8]. The authors consider semantic heterogeneity and URI-based entity identification solved by a semantic entity resolution method based on inference mechanism using rules to manage the misunderstanding of data. A formal knowledge structure based on a conceptualization represented by semantic web approach [48], is a basic approach to reduce *Variety* issue and align different heterogeneous repositories [42] in complex scenario as the financial domain. In this context the use of Big Data sanctification give us new tools for classifying information based on specific knowledge structures. The MOUNT system is presented in [43]. It aims at representing and processing a large-scale heterogeneous big data generated from multiple sources using multi-level semantic annotation and query processing.

Focusing on the specific task related to document visualization using tag clouds, our approach uses a keywords extraction technique to construct a summary through tag clouds. The quality of the extracted keywords depends on the algorithms used and various methods have been proposed in the literature [51]. Artificial intelligence techniques in addition to semantics [30] and co-occurrence [37] were extensively used for extracting keywords from a single document. In [4] different information is used based on the VSM and genetic algorithms to compute a grade of similarity between sentences and weights associated to the features. In [29] the authors use linguistic features to represent the relevance of the term, also based on the position in the document. In [53], a tag-oriented approach based on a linear transformation to measure the tag relevance is discussed. Cluster-based algorithms have also been proposed, for example in [28] k-means has been used to semantically group similar tags. Another interesting method based on the co-occurrences of tags with the aim of comparing the structure of the folksonomies network is proposed by [20] in which the authors have analyzed the similarities between the tags and the documents to enrich semantic and hierarchical aspects. In [38] a model for the union of ontologies and social networks with the use of tags mechanisms has been presented. In [21] the authors present a technique to support collaborative semantic understanding of generated tags. Their approach gathers tags in different semantic groups and the relations among tags is expressed through the visual distance between them.

The proposed strategy presents several novelties. Our framework offers a high level of generalization compared with the presented literature where the discussed approaches and frameworks face the issue of the *Variety* dimension in specific domains or without considering multimedia information. Following our approach, we design and implement a formal semantic-based model to fully represent and manage specific and general knowledge domains. Moreover, we consider some standard multimedia descriptors to give a possible solution to the question of heterogeneity. In addition, we propose a different strategy based on the dynamic extraction of networks from our SMBD and use a metric for the generation of tags based on a combination of textual and visual analysis.

3 The Semantic Multimedia Big Data

The SMBD implemented in this paper is based on a model proposed in [49]. A general concept can be represented in various multimedia forms (i.e. signs) 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 model structure consists of a triple $< \mathbf{S}, \mathbf{P}, \mathbf{C} >$ defined as: (i) S: the set of signs; (ii) P: the set of properties used to relate signs with concepts; (iii) C: the set of constraints on the P collection.

We use two types of representations (MM): *word* (i.e. text), and the *visual* (i.e. images). The properties are linguistic relations and the constraints contain validity rules applied to properties with respect to the considered multimedia. Knowledge is conceptual represented by an ontology and by a *Semantic Network* (SN) from a logic point of view. It is a graph structure where nodes are concepts and arcs are linguistic relation between them. The concept is a set of multimedia data representing an abstract idea. The language chosen to describe this model is the DL version of OWL. It consents the declaration of disjointed classes to state, for example, that a word can belong to a syntactic category. It is also possible to declare union classes used to specify domains, ranges and properties to relate concepts and multimedia nodes. The connections in the semantic network are represented as *ObjectProperties* and have constraints that depend on the syntactic category or type of property (semantic or lexical). Some examples are described in the Table 1.

As examples of constraints, the hyperonomy property can only be used between nouns and nouns or between verbs and verbs. Each multimedia is linked to the represented concept by the ObjectProperty *hasConcept* and vice versa with *hasMM*. They are the only properties that can be used to link concepts with multimedia. The other properties are used to link multimedia to multimedia or concept to concepts. The attributes of Concept and Multimedia classes are also described. The concept has as attribute: *Name* which represents the name of the concept and the field *Glossary* which contains a short description of it. The common attributes of the MM subclasses are *Name* and *ID*. Each subclass has its own set of features depending on the nature of the media. In the visual case we use the PHOG [9] global feature descriptor which has good

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

Table 2. Properties constraints

Constraint	Class	Property	Constraint range
AllValuesFrom	Noun	Hypernym	Noun
AllValuesFrom	Adjective	attribute	Noun
AllValuesFrom	Verb	also see	Verb
AllValuesFrom	Noun	Hyponym	Noun

performance with respect to other descriptors [41]. This descriptor consists of a histogram of orientation gradients over each image subregion at each resolution level. The distance between two PHOG image descriptors reflects the property of images to contain similar shapes in corresponded spatial layout. The use of union classes simplify the domain rules but at the same time the model does not exhibit perfect behavior. For example, the property of hyperonymy allows the relationship between nouns and verbs. In this context, we have to define different constraints to represent how the linguistic properties relate concepts and/or MM. Table 2 shows some of the these constraints.

In some cases, the existence of a property between two or more individuals involves the existence of other properties. For example, if a concept A is a hyponym of a concept B, the concept B is hyperonym for A. These features are described using properties features and some examples are in Table 3.

Our model and the related SN have been implemented in a SMBD using the*Neo4j* graph database and it has been populated using WordNet [39] and ImageNet [24]. Therefore, we can consider the same knowledge organization of these sources. The Fig. 1 is a macro visualization the resulting graph limited to 24271 nodes and 40000 relationships [10,11].

The goal of this figure is to show the complexity of our SMBD.

4 The Visualization Framework

The visualization of the document summarization in described and discussed in the following of this section. It uses a Semantic Tag cloud where the task of

Table 3. Properties features

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

tag recognition (i.e. the extracted keywords) is based on a Word Sense Disambiguation (WSD) step. The WSD is a basic action because the property of a term to express different meaning; this linguistic property is called polysemy. The implemented knowledge base (i.e. SMDB) is the used as support for the visualization and the WSD tasks and also to add semantic related terms to the recognized tags. We use our SMBD in the visualization process and both in the WSD step and to retrieve additional keywords. The system executes the WSD task to assign the right sense to each term. This step is based on the analysis of the term context. Therefore, each term sense is compared with all the senses of the other terms in a document. We use a semantic based similarity metric to measure the relatedness between terms and we assign to the considered term the sense with the high similarity score value. First of all we assign to the defined linguistic properties a weight σ_i, which represents the straighten of each relation from a expressive power point of view. In our opinion the linguistic relations have different power when they connect concepts or words. Also other authors support our intuition [19]. The weight assigned to each property is set following the measures defined in the work cited just now and we extend them to similar properties. They are real numbers in the $[0, 1]$ interval. The defined metric is composed of two components: the path length (l) between pairs of terms and the depth (d) of their subsumer, represented by the number of hops. This correlation is the semantic relatedness between the considered terms and it is computed through a nonlinear function. The use of this kind of function depends to different questions. The score of path length and depth may range from 0 to infinity due to their definition. On the other hand, the terms relatedness is represented with a number in the $[0, 1]$ interval. When the path length decreases toward 0, the relatedness should monotonically increase toward 1, on the contrary it should monotonically decrease toward 0 when path length goes to infinity. Moreover we have to consider a scaling effect regarding the depth, because a word in a high semantic hierarchy level expresses a more general concept than one in a lower level. A non linear function is able to scale down the contribution of subsumers in a upper level and scale up those in a lower one.

We are now in the position of present the metric components and later the our similarity function.

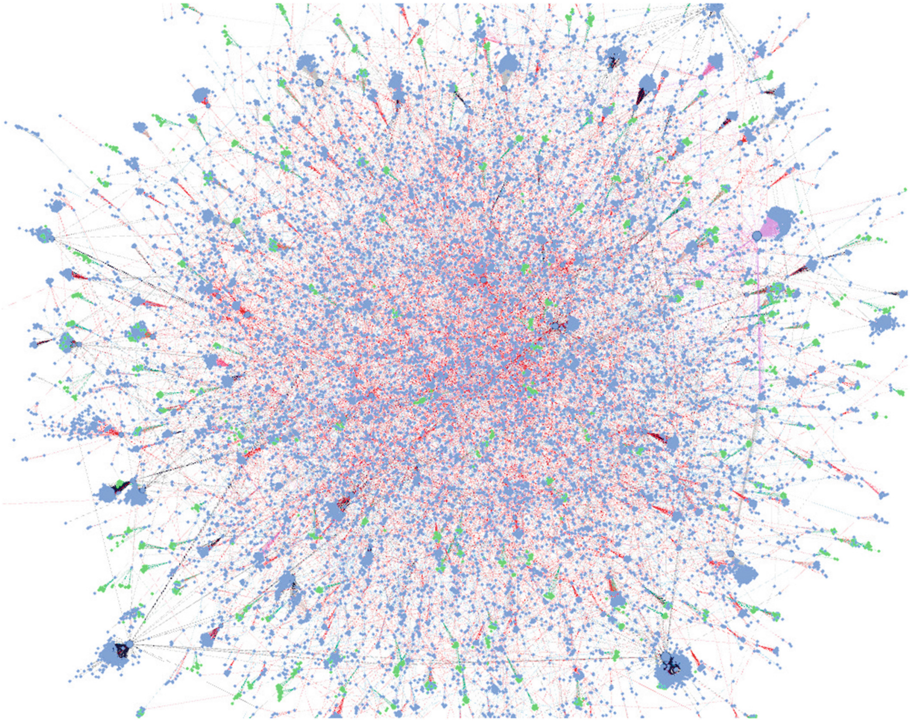


Fig. 1. Macro visualization of Neo4j db with Gephi

Given two words w_1 and w_2, the length l of the path between w_1 and w_2 is computed using our SMBD and it is defined as:

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

where j spans over all the paths between w_1 and w_2, $h_j(w_1, w_2)$ is the number of hops in the j-th path and σ_i is the weight assigned to the i-th hop in the j-th path in respect to the hop linguistic property.

We use our SMDB to measure the depth d of w_1 and w_2 subsumer. Te compute this measure we consider only the IS-A hierarchy (i.e. the hyponymy and hyperonymy structure). Moreover, the depth is measured as the number of hops from the subsumer of w_1 and w_2 to the root of the hierarchy. With this aim, we use an exponential function to solve the issues discussed above.

We define the *Semantic Similarity Score* (SSS) as:

$$SSS(\upsilon) = \sum_{(w_p} e^{-\alpha \cdot l(w_p)} \frac{e^{\beta \cdot d(w_p)} - e^{-\beta \cdot d(w_p)}}{e^{\beta \cdot d(w_p)} + e^{-\beta \cdot d(w_p)}} \tag{2}$$

where (w_p) are all pairs of words in the document v and $\alpha \geq 0$ and $\beta > 0$ are two scaling factors (i.e. $\alpha = 0.2$ and $\beta = 0.6$) [32].

The proposed function is used to compute the WSD for all the senses related to the extracted term. The WSD task is performed in a window of context and in our approach we use the entire document. The sense with the maximum score computed by our metric is associated as right sens to the considered term. This task returns a list of concepts represented with their terms and ordered by the SSS. Moreover, we use also visual information in the WSD task. The images are taken from the analyzed document to measure the similarity with respect to the our SMBD visual collection. An image is represented by the same global feature descriptor used in our big data (i.e. PHOG). This feature is compared with the ones in the SMBD through a distance metric which measures the proximity between two points belonging to the same metric space (i.e. the feature space). Since every visual node in the database is linked to a concept, the result is a list of concepts, sorted by the evaluated metric. We use the cosine metric to measure the similarity. In the whole WSD process we consider as good candidates the concepts in common between the textual and visual WSD. The combination of textual and visual analysis is combined using the SUM function [31] and the chosen sense represented by term/image is the one with the highest value.

The Visual Semantic Tag Cloud is built following a novel technique. It is based on a combination of term semantic properties and statistical information. We use the weight associated to the recognized keywords and their relations with the same terms in our knowledge base The importance of each term is measured considering its polisemy grade. We argue that the polysemy expresses the ambiguity in the use of a terms when it can have multiple meanings. This value is calculated using our knowledge base.

We called this measure *centrality* and it is calculate for a generic term i as:

$$\chi_i = \frac{1}{poly_i} \tag{3}$$

$poly_i$ is the associated number of senses of i.

Querying the knowledge base the word *lion* has, for example, four associated senses therefore its chance to represent a specific sense is equal to 1/4.

The proposed metric uses statistic information based on term-weight function to better represent relevant document terms. Our approach takes into consideration two different normalized components [50]:

- *Term frequency*- the number of occurrences of a term in a document;
- *Document size factor*- it compensates for high term frequencies of terms in large documents.

We are now in the position of presenting our metric to visualize the summarized document:

$$V_{i,k} = \frac{(a + (1-a)(TF_{i,k}/TF_{max,k}))\chi_i}{\sqrt{\sum_{i \in k}(a + (1-a)(TF_{i,k}/TF_{max,k})(\chi_i))^2}} \tag{4}$$

where k is the set of terms representing to the j-th document, i is the i-th term, $TF_{i,k}$ represents the term frequency of i in k, $TF_{max,k}$ being the highest value of the term frequency in k, ϖ_i is the centrality of i, a is used as a smoothing factor to dump TF by the largest TF value in k related to the second metric component. The need is to not allow important changing of the normalized $TF_{i,k}$ for small variations of $TF_{i,k}$. The suggested line in literature is to set a=0.5 [36].

The proposed metric add statistical measures to efficiently visualize document summary also combining semantic information as centrality. In this way we can have an improving in the recognition of relevant terms.

These information are used to compute the font size in the visualization of the semantic tag cloud. Moreover, we add semantic related terms through our knowledge base. This information enrichment is performed by a semantic network extraction created by the recognized terms in the WSD step. Starting from the concepts represented by these terms we add other direct linked concepts considering all linguistic relations counting out hyperonym to avoid general concepts. These terms are visualize by the same original term font size with a scaling factor computed by Eq. 2.

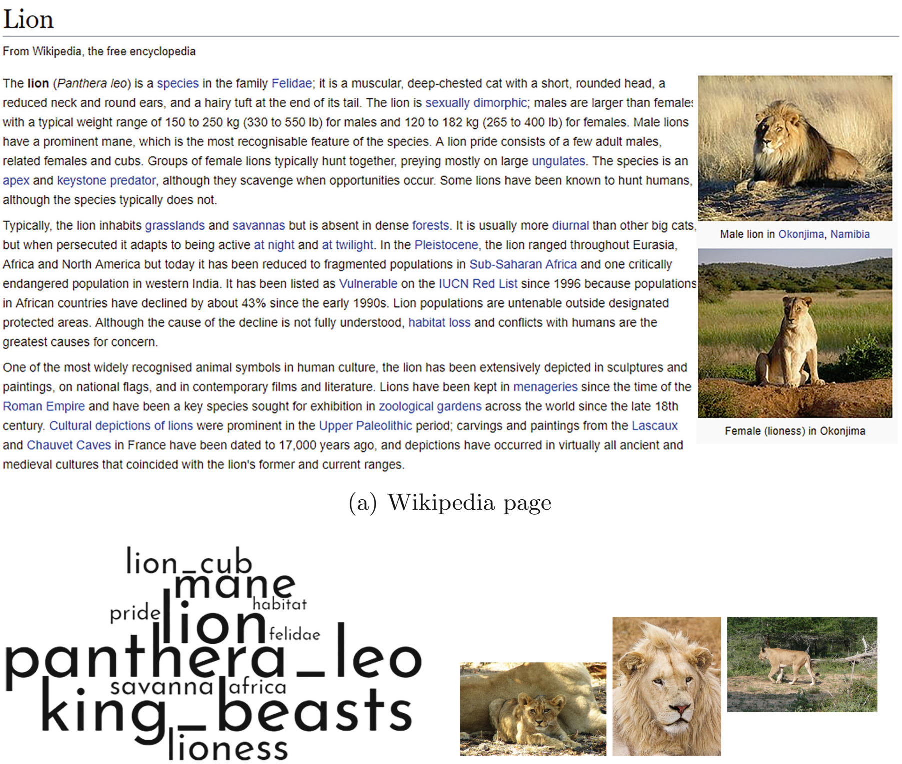
In the following section we will present some experiments to show that the tag cloud enrichment with new terms semantically related to the ones extracted from the analyzed document but not present in it improves the perception of the document topic by users.

5 Case Study and Evaluation

This section is focused on the implementation of a Semantic Tag Cloud for document visualization. We describe the procedure for the generation of our document summary visualization tool considering a real document from Wikipedia in a general conceptual domain about *animal* and about a specific topic i.e. *lion* (http://en.wikipedia.org/wiki/Lion).

The web document has been preprocessed with the aim of transform in a basic form all the extracted terms. The WSD task is later performed to disambiguate the analyzed terms and have their right sense. The correct meaning is recognized using the process described in the previous section by means of visual and textual information measuring the semantic relatedness between all keywords in the document and visual features. An example of disambiguated term is "*lion*".; we extract four senses from our knowledge base derived from the integration of WordNet and ImageNet in our SMBD. The glossary of these senses are:

1. lion, king of beasts, Panthera leo: large gregarious predatory feline of Africa and India having a tawny coat with a shaggy mane in the male;
2. lion, social lion: a celebrity who is lionized (much sought after);
3. Leo, Lion: (astrology) a person who is born while the sun is in Leo;
4. Leo, Leo the Lion, Lion: the fifth sign of the zodiac; the sun is in this sign from about July 23 to August 22.



Lion

From Wikipedia, the free encyclopedia

The **lion** (*Panthera leo*) is a species in the family Felidae; it is a muscular, deep-chested cat with a short, rounded head, a reduced neck and round ears, and a hairy tuft at the end of its tail. The lion is sexually dimorphic; males are larger than females with a typical weight range of 150 to 250 kg (330 to 550 lb) for males and 120 to 182 kg (265 to 400 lb) for females. Male lions have a prominent mane, which is the most recognisable feature of the species. A lion pride consists of a few adult males, related females and cubs. Groups of female lions typically hunt together, preying mostly on large ungulates. The species is an apex and keystone predator, although they scavenge when opportunities occur. Some lions have been known to hunt humans, although the species typically does not.

Typically, the lion inhabits grasslands and savannas but is absent in dense forests. It is usually more diurnal than other big cats, but when persecuted it adapts to being active at night and at twilight. In the Pleistocene, the lion ranged throughout Eurasia, Africa and North America but today it has been reduced to fragmented populations in Sub-Saharan Africa and one critically endangered population in western India. It has been listed as Vulnerable on the IUCN Red List since 1996 because populations in African countries have declined by about 43% since the early 1990s. Lion populations are untenable outside designated protected areas. Although the cause of the decline is not fully understood, habitat loss and conflicts with humans are the greatest causes for concern.

One of the most widely recognised animal symbols in human culture, the lion has been extensively depicted in sculptures and paintings, on national flags, and in contemporary films and literature. Lions have been kept in menageries since the time of the Roman Empire and have been a key species sought for exhibition in zoological gardens across the world since the late 18th century. Cultural depictions of lions were prominent in the Upper Paleolithic period; carvings and paintings from the Lascaux and Chauvet Caves in France have been dated to 17,000 years ago, and depictions have occurred in virtually all ancient and medieval cultures that coincided with the lion's former and current ranges.

Male lion in Okonjima, Namibia

Female (lioness) in Okonjima

(a) Wikipedia page

(b) Text Semantic Cloud

(c) Image Semantic Tag Cloud

Fig. 2. Semantic Tag Cloud Example

Only the first one has a high similarity with all the other word of analyzed terms and the images in the original document. The list of disambiguated terms has benn used to build a SN for each term and, following the strategy described before new terms have been found to enrich our visualization. A complete example of the Semantic Cloud generation process is in Fig. 2.

The number of shown terms and images can be set by the user. The semantic tag cloud shows the terms and images more similar to the representative concept for the analyzed document. On the other hand, the use of a semantic big data give us also a visual representation of the considered concept. Therefore, the visual cloud has been generated using the images related to the concepts recognized in the WSD. We explicit point out that some concepts haven't related images due to the knowledge source (i.e. ImageNet) used to populate our SMBD.

We have to highlight the improvements of the presented visualization approach with a proper evaluation of the proposed techniques. At this stage of our research, we want to measure the user perception in the using of our visualization tool. We are interested in this kind of evaluation due to intrinsic subjective nature of the understanding process of document topic. For this reason, we need to take into account specific indicators to measure the user perception.

We chose to use a general and well-known knowledge source as Wikipedia and fetch 500 web pages related to animal domain and generate the related summaries represented by semantic clouds. Therefore, they are analyzed by a questionnaire asked to a group of 100 users (MSc students and Ph.D. students of information science). In our experiments we present a comparison of our textual (**VC**) and visual (**VC**) techniques with common document term frequency counting (**TC**) for tag cloud generation.

Our approach has been evaluated using a methodological framework discussed in [40]. This methodology is devoted to the service evaluation and it is based on two different kinds of statements:

- **E** - a general evaluation on service category;
- **P** - an evaluation of a particular service.

A 7-point scale in a *strongly disagree* (**1**) to *strongly agree* (**7**) interval is used to measure each statement. Moreover, we evaluate our strategy using others measures [22,33] to put into account different features of our methodology:

- **PU** - Perceived Usefulness;
- **PEU** - Perceived Ease of Use;
- **PE** - Perceived Enjoyment.

The document test set and the related tag clouds have been randomly assigned to users and evaluated using these indicators.

The results are shown in Fig. 3 in terms of mean.

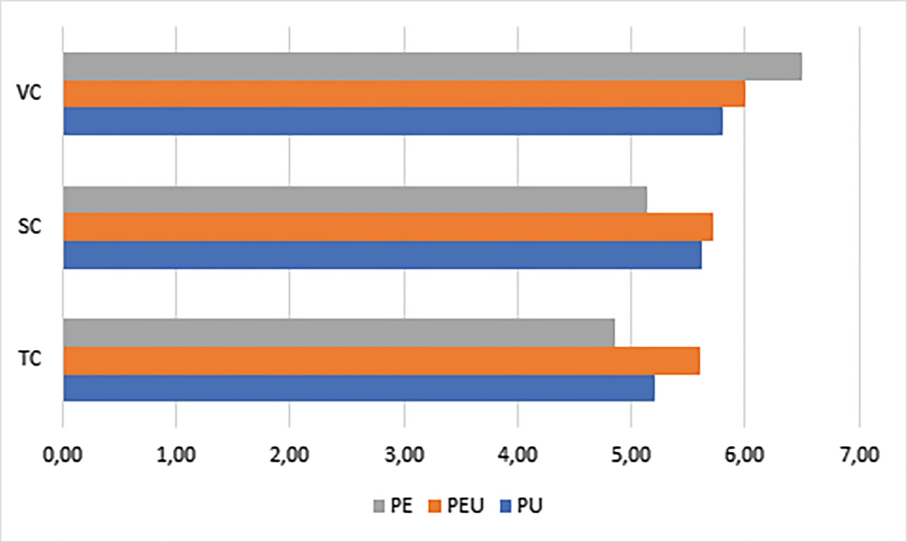


Fig. 3. Experimental results

The proposed textual and visual tag clouds visualization strategies shown better results in terms of topic document understanding compared with the standard frequency-based tag cloud. Moreover, the enrichment of our visualization structure with terms semantically related to the ones extracted from the analyzed documents gives an improvement of the user knowledge about the document topic.

6 Conclusion

Efficient approaches to analyze documents and represent their contents in an improved view can provide a solution for information overload. In this context the integration of semantic web vision and big data technologies represent an interesting way to organize and store data. Semantic analysis techniques and document visualization tools offer a concise way to represent huge volume of data and support people to recognize and understand information from the original sources [12,14]. In this paper a semantic multimedia big data has been presented. Information has been organized and stored using a multimedia model in which the concepts are represented in textual and visual form. The model has been implemented using a graph big data. Moreover, a strategy based on semantic tag cloud has been proposed to visualize key concepts related to analyzed documents A real case study on the semantic tag cloud generation and visualization has been presented to show the effectiveness of our approach. We are considering to extend our framework with other statistical components and implement novel browsing strategies by means of textual and visual tags to enhance basic tasks of the information retrieval process [2]. In addition, we are investigating on the use of other visual description based on deep neural networks to improve the accuracy of our strategy and perform quantitative analysis of our approach compared with similar base lines and in the integration of domain ontologies in our knowledge base to improve its topic coverage and multimedia representations [13,15,47].

References

1. Agerri, R., Artola, X., Beloki, Z., Rigau, G., Soroa, A.: Big data for natural language processing: a streaming approach. Knowl.-Based Syst. **79**, 36–42 (2015)
2. Albanese, M., Capasso, P., Picariello, A., Rinaldi, A.: Information retrieval from the web: an interactive paradigm. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3665, LNCS, pp. 17–32 (2005)
3. Albanese, M., Maresca, P., Picariello, A., Rinaldi, A.: Towards a multimedia ontology system: an approach using tao xml. In: Proceedings: DMS 2005–11th International Conference on Distributed Multimedia Systems, pp. 52–57 (2005)
4. Alguliev, R.M., Aliguliyev, R.M.: Effective summarization method of text documents. In: Proceedings The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005, pp. 264–271. IEEE (2005)
5. Bansal, S.K., Kagemann, S.: Integrating big data: a semantic extract-transform-load framework. Computer **48**(3), 42–50 (2015)

6. Begelman, G., Keller, P., Smadja, F., et al.: Automated tag clustering: improving search and exploration in the tag space. In: Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, pp. 15–33 (2006)
7. Bello-Orgaz, G., Jung, J.J., Camacho, D.: Social big data: recent achievements and new challenges. Inf. Fusion **28**, 45–59 (2016)
8. Benbernou, S., Huang, X., Ouziri, M.: Semantic-based and entity-resolution fusion to enhance quality of big RDF data. IEEE Transactions on Big Data (2017)
9. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, pp. 401–408. ACM (2007)
10. Caldarola, E., Picariello, A., Rinaldi, A.: Big graph-based data visualization experiences: the wordnet case study. In: IC3K 2015 - Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 1, pp. 104–115 (2015)
11. Caldarola, E., Picariello, A., Rinaldi, A.: Experiences in wordnet visualization with labeled graph databases. Commun. Comput. Inf. Sci. **631**, 80–99 (2016)
12. Caldarola, E., Picariello, A., Rinaldi, A., Sacco, M.: Exploration and visualization of big graphs the dbpedia case study. In: IC3K 2016 - Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 1, pp. 257–264 (2016)
13. Caldarola, E., Rinaldi, A.: An approach to ontology integration for ontology reuse. In: Proceedings - 2016 IEEE 17th International Conference on Information Reuse and Integration, IRI 2016, pp. 384–393 (2016)
14. Caldarola, E., Rinaldi, A.: Improving the visualization of word net large lexical database through semantic tag clouds. In: Proceedings - 2016 IEEE International Congress on Big Data, BigData Congress 2016, pp. 34–41 (2016)
15. Caldarola, E., Rinaldi, A.: A multi-strategy approach for ontology reuse through matching and integration techniques. Adv. Intell. Syst. Comput. **561**, 63–90 (2018)
16. Caldarola, E.G., Rinaldi, A.M.: Big data: a survey: the new paradigms, methodologies and tools. In: DATA 2015–4th International Conference on Data Management Technologies and Applications, Proceedings, pp. 362–370 (2015)
17. Caldarola, E.G., Rinaldi, A.M.: Big data visualization tools: a survey: the new paradigms, methodologies and tools for large data sets visualization. In: DATA 2017 - Proceedings of the 6th International Conference on Data Science, Technology and Applications, pp. 296–305 (2017)
18. Caldarola, E.G., Rinaldi, A.M.: Modelling multimedia social networks using semantically labelled graphs. In: 2017 IEEE International Conference on Information Reuse and Integration (IRI), pp. 493–500 (2017)
19. Castano, S., Ferrara, A., Montanelli, S.: H-match: an algorithm for dynamically matching ontologies in peer-based systems. In: In Proceedings of WebS, pp. 231–250 (2003)
20. Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V.D., Loreto, V., Hotho, A., Grahl, M., Stumme, G.: Network properties of folksonomies. Ai Commun. **20**(4), 245–262 (2007)
21. Chen, Y.-X., Santamaría, R., Butz, A., Therón, R.: TagClusters: semantic aggregation of collaborative tags beyond tagclouds. In: Butz, A., Fisher, B., Christie, M., Krüger, A., Olivier, P., Therón, R. (eds.) SG 2009. LNCS, vol. 5531, pp. 56–67. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02115-2_5
22. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**(3), 319–340 (1989)

23. De Mauro, A., Greco, M., Grimaldi, M.: A formal definition of big data based on its essential features. Library Rev. **65**(3), 122–135 (2016)
24. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 248–255. IEEE (2009)
25. Emani, C.K., Cullot, N., Nicolle, C.: Understandable big data: a survey. Comput. Sci. Rev. **17**, 70–81 (2015)
26. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**(2), 199–220 (1993)
27. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social bookmarking tools (i) a general review. D-lib Magazine **2**(4) (2005). http://www.dlib.org/dlib/april05/hammond/04hammond.html
28. Hassan-Montero, Y., Herrero-Solana, V.: Improving tag-clouds as visual information retrieval interfaces. In: International Conference on Multidisciplinary Information Sciences and Technologies, pp. 25–28. Citeseer (2006)
29. Hu, X., Wu, B.: Automatic keyword extraction using linguistic features. In: Sixth IEEE International Conference on Data Mining Workshops, 2006, ICDM Workshops 2006, pp. 19–23. IEEE (2006)
30. 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)
31. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
32. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng. **15**(4), 871–882 (2003)
33. Lin, C.S., Wu, S., Tsai, R.J.: Integrating perceived playfulness into expectation-confirmation model for web portal context. I&M **42**(5), 683–693 (2005)
34. Lv, Z., Song, H., Basanta-Val, P., Steed, A., Jo, M.: Next-generation big data analytics: state of the art, challenges, and future research topics. IEEE Trans. Ind. Inform. **13**(4), 1891–1899 (2017)
35. Mami, M.N., Scerri, S., Auer, S., Vidal, M.-E.: Towards semantification of big data technology. In: Madria, S., Hara, T. (eds.) DaWaK 2016. LNCS, vol. 9829, pp. 376–390. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43946-4_25
36. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
37. 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)
38. Mika, P.: Ontologies are us: a unified model of social networks and semantics. Web Semant.: Sci., Serv. Agents World Wide Web **5**(1), 5–15 (2007)
39. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
40. Parasuraman, A., Zeithaml, V.A., Berry, L.L.: Servqual: a multiple-item scale for measuring consumer perceptions of service quality. J. Retailing **64**(1), 12–40 (1988)
41. Purificato, E., Rinaldi, A.M.: Multimedia and geographic data integration for cultural heritage information retrieval. Multimedia Tools Appl. **77**, 1–23 (2018)
42. Quboa, Q., Mehandjiev, N.: Creating intelligent business systems by utilising big data and semantics. In: 2017 IEEE 19th Conference on Business Informatics (CBI), vol. 2, pp. 39–46. IEEE (2017)

43. Rani, P.S., Suresh, R.M., Sethukarasi, R.: Multi-level semantic annotation and unified data integration using semantic web ontology in big data processing. In: Cluster Computing (2017)
44. Rinaldi, A.: Improving tag clouds with ontologies and semantics. In: Proceedings - International Workshop on Database and Expert Systems Applications, DEXA, pp. 139–143 (2012)
45. Rinaldi, A.: Document summarization using semantic clouds. In: Proceedings - 2013 IEEE 7th International Conference on Semantic Computing, ICSC 2013, pp. 100–103 (2013)
46. Rinaldi, A.: Web summarization and browsing through semantic tag clouds. Int. J. Intell. Inf. Technol. **15**(3), 1–23 (2019)
47. Rinaldi, A., Russo, C.: A matching framework for multimedia data integration using semantics and ontologies. In: Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018, vol. 2018, pp. 363–368 (2018)
48. Rinaldi, A., Russo, C.: A semantic-based model to represent multimedia big data. In: MEDES 2018–10th International Conference on Management of Digital EcoSystems, pp. 31–38 (2018)
49. Rinaldi, A.M.: A multimedia ontology model based on linguistic properties and audio-visual features. Inf. Sci. **277**, 234–246 (2014)
50. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. In: Sparck Jones, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 323–328. Morgan Kaufmann Publishers Inc., San Francisco (1997)
51. Siddiqi, S., Sharan, A.: Keyword and keyphrase extraction techniques: aliterature review. Int. J. Comput. Appl. **109**(2), 18–23 (2015)
52. Xu, Z., Wei, X., Luo, X., Liu, Y., Mei, L., Hu, C., Chen, L.: Knowle: a semantic link network based system for organizing large scale online news events. Fut. Gener. Comput. Syst. **43**, 40–50 (2015)
53. Zhu, J., et al.: Tag-oriented document summarization. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1195–1196. ACM (2009)