A Novel Framework to Represent Documents using a Semantically-grounded Graph Model

Antonio M. Rinaldi[1] and Cristiano Russo[2]
[1]*Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, IKNOS-LAB-Intelligent and Knowledge Systems-LUPT, University of Naples Federico II, Italy*
[2]*LISSI Laboratory, University of Paris-Est Creteil (UPEC), France*

Keywords: Document Representation, Semantic and Linguistic Analysis, WordNet, Lexical Chains, NoSQL, Neo4J.

Abstract: As an increasing number of text-based documents, whose complexity increases in turn, are available over the Internet, it becomes obvious that handling such documents as they are, i.e. in their original natural-language based format, represents a daunting task to face up for computers. Thus, some methods and techniques have been used and refined, throughout the last decades, in order to transform the digital documents from the full text version to another suitable representation, making them easier to handle and thus helping users in getting the right information with a reduced algorithmic complexity. One of the most spread solution in document representation and retrieval has consisted in transforming the full text version into a vector, which describes the contents of the document in terms of occurrences patterns of words. Although the wide adoption of this technique, some remarkable drawbacks have been soon pointed out from the researchers' community, mainly focused on the lack of *semantics* for the associated terms. In this work, we use WordNet as a generalist linguistic database in order to enrich, at a semantic level, the document representation by exploiting a label and properties based graph model, implemented in Neo4J. This work demonstrates how such representation allows users to quickly recognize the document topics and lays the foundations for cross-document relatedness measures that go beyond the mere word-centric approach.

1 INTRODUCTION

Text is the most traditional method for information recording and knowledge representation (Yan and Jin, 2012) and a common source of learning in an instructional setting (Thorndyke, 1978). Actually, humans do not judge text relatedness merely at the level of text words, since, words trigger reasoning at a much deeper level, i.e., that of *concepts* - the basic units of meaning that serve humans to organize and share their knowledge. Humans interpret the specific wording of a document in the much larger context of their background knowledge and experience (Gabrilovich and Markovitch, 2007). While reasoning about semantic relatedness of natural language utterances is routinely performed by humans, it remains an insurmountable obstacle for computers. Thus, some methods and techniques have been used and refined, throughout the last decades, in order to transform the digital document from the full text version to another suitable representation, making them easier to handle for automated software agents in Information Retrieval (IR) Systems. IR-models are based on strategies that span from the set-theoretical boolean methods for IR to the algebraic Space Model Vector and the Latent Semantic Indexing and, finally, to the Topic-based Space Vector Model. Particularly, ideas underlined by the Space Vector Model are the most spread solution in document representation and retrieval and consist in transforming the full text version into a vector which describes the contents of the document in terms of occurrences patterns of words. Although the wide adoption of this technique, some remarkable drawbacks have been soon pointed out from the researchers community, mainly focused on the lack of *semantics* for the associated terms, namely, it does not say anything about the nature of the meaning of terms pairs (e.g., if they are synonyms or linked somehow at a semantic-level). Consequently, new strategies making use of external linguistic resources have been increasingly adopted to imbue words with semantics and linking terms together with linguistic-semantic relations (Rinaldi, 2009) and very large knowledge base representation (Caldarola et al., 2015). In our

Rinaldi, A. and Russo, C.
A Novel Framework to Represent Documents using a Semantically-grounded Graph Model.
DOI: 10.5220/0006932502030211
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR*, pages 203-211
ISBN: 978-989-758-330-8

context, one of the most spread linguistic resources are lexical databases like WordNet (Miller, 1995). We use WordNet as a generalist linguistic database, in order to semantically augment the document representation going beyond the mere word-centric approach. This way, we try to demonstrate how, just by analysing the topology of the expanded lexical chains, represented through a labelled-based graph model, it is possible to predict the knowledge categories the document belongs to, regardless any statistical measures related to the document terms. Our solution has been implemented in a newly adopted tool from the NoSQL technologies, namely, Neo4J, which allows us to represent the expanded lexical chains through a properties and label-based graph model, able to be horizontally scaled and distributed. This work focuses on document representation/visualization exploiting the features available from the new tool mentioned above.

The reminder of the paper is structured as follows. After a state-of-the-art of the main issues and solutions regarding the document representation for text-mining and retrieval, provided in section 2, an overview of the system architecture for our document representation solution is presented in section 3, along with the implementation details. Section 4 motivates the reasons for using the properties and labels-based graph model in order to represent the document, outlining the procedure for obtaining the graph representation and motivating the choice for the selected document corpus. Section 5 shows the results of applying the solution over some documents, as examples, while section 6 draws the conclusion outlining the major findings and laying the foundations for future investigations.

2 RELATED WORKS

In this section, we present and discuss a literature focused on document visualization and text categorization techniques. In this regard, several methods have been proposed to help users in searching, visualizing and retrieving useful information from a text-based corpora. One of these approaches is based on the creation of tag clouds, which can be used for basic user-centered tasks (Rivadeneira et al., 2007). Other studies improve tag cloud statistical based approach with semantic information (Rinaldi, 2012; Rinaldi, 2013). In our approach, we use a keywords extraction technique to build the semantically-expanded lexical chain. The quality of extracted keywords depends on the corresponding keyword extraction algorithm and several methods have been proposed in the literature. In (Hu and Wu, 2006) the authors use linguistic features to represent the importance of the word position in a document. They extract topical terms and their previous-term and next-term co-occurrence collections using several methods. A tag-oriented summarization approach is discussed in (Zhu et al., 2009). The authors present a new algorithm using a linear transformation to estimate the importance of tags. The tags are further expanded to include related words using association mining techniques. The final summary is generated with a sentence evaluation based on expanded tags and TF-IDF of each word in a sentence. An iterative approach for document keywords extraction based on the relationship between different granularities (i.e., relationships between words, sentences, and topics) is presented in (Wei, 2012). The method is first implemented by constructing a graph, which reflects relationships between different size of granularity nodes, and then using an iterative algorithm to calculate score of keywords; the words with highest score are chosen as keywords. In (Kaptein, 2012) the author describes an application in which word clouds are used to navigate and summarize Twitter search results. This application summarizes sets of tweets into word clouds, which can be used to get a first idea of the contents of the tweets. Moreover, several studies have been presented to add more information to folksonomies and enhance tag visualization in order to improve the use of tag clouds. Several approaches (Begelman et al., 2006; Fujimura et al., 2008) have been proposed to measure tag similarity using statistics. Clustering algorithms were applied to gather semantically similar tags. In (Hassan-Montero and Herrero-Solana, 2006) the k-means algorithm was applied to group semantically similar tags. Li et al. (Li et al., 2007) supported a large scale social annotations browsing based on an analysis of semantic and hierarchical relations. An approach to build semantic networks on the basis of tag co-occurrences and network structures of folksonomies is in (Cattuto et al., 2007). The same authors analyzed similarities between tags and documents in order to enrich semantic aspects of social tagging. An interface for information searching task using tag clouds has been presented in (Sinclair and Cardew-Hall, 2008). The authors point out that tag clouds satisfy all the roles mentioned in (Rivadeneira et al., 2007), as visual summaries of content, and they observed that the process of scanning the cloud and clicking on tags is easier than the formulation of a search query. In (Chen et al., 2009) the authors investigate ways to support semantic understanding of collaboratively generated tags. They conducted a survey on practical tag usage in Last.fm.

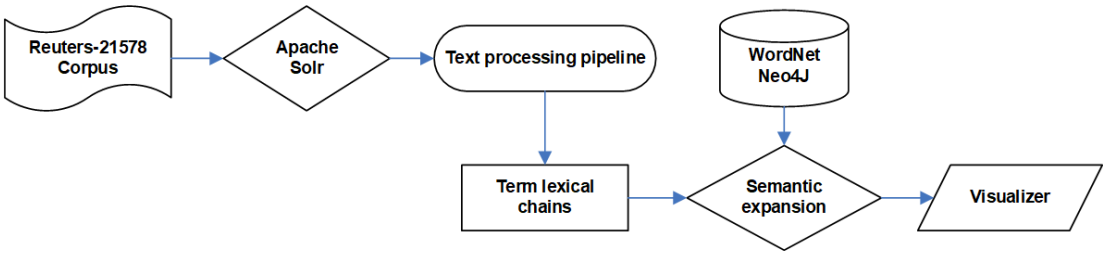

Figure 1: An high-level view of the document representation system architecture.

3 DOCUMENT REPRESENTATION SYSTEM ARCHITECTURE

Figure 1 shows a high-level view of the implemented system. The main blocks depicted in the figure are: the Apache Solr search server containing and indexing the Reuters-21578 corpus (Lewis, 1997) used as document collection, the text-processing pipeline, which contains also the tokenizer, which obtains a normalized lexical chain for each document and, finally, the semantic expansion block, which makes use of Neo4J and WordNet to represent each document as an expanded and semantically-grounded lexical chain. Apache Solr is an open-source search platform built on Apache Lucene (McCandless et al., 2010) allowing the storage and the indexing of large volume of documents collections. Solr allows a data-driven schemaless mode when populating the index but also a fine-grain control over the schema production time. In this work, the Solr Java APIs have been used in order to read the Reuters documents and store them in a coherent and searchable document index. Reuters collection is a resource for research in information retrieval, machine learning, and other corpus-based research. Documents were marked up with SGML tags, and a corresponding SGML DTD was produced, so that the boundaries of important sections of documents (e.g., category fields) are unambiguous. Each article has a structure which highlight the fields used in this work as: *Topic* attribute of the Reuters main tag, which is a boolean flag indicating if the document has been categorized by human indexers, i.e., if the document is in the training set; *NEWID* another attribute of the main tag, which assigns a unique ID to each doc in chronological order; *TOPICS*, which encloses the list of TOPICS categories, if any, for the document. The other fields, i.e., PLACES, ORGS, EXCHANGES and COMPANIES are same as TOPICS but for the corresponding typology of categories. In this work we use only the TOPICS categorization; *AUTHOR*, the author of the story; *TITLE*, the title of the story; *BODY*, The main text of the story. It may has a normal structure or can be a brief text containing one or at most two lines. In this work, we consider only normal type document. A test collection for text categorization contains a set of texts and, for each text, a specification of what categories that text belongs to. For the Reuters-21578 collection the documents are Reuters newswire stories, and the categories are five different sets of content related categories. The TOPICS categories are economic subject categories, e.g., ”gold”, ”inventories”, and ”money-supply”. As described in section 5, the proposed methodology has been applied to one hundred documents coming from ten different categories. Thanks to the search capabilities of Solr and the schema-based representation of the document as a set of fields has been possible to quickly harvest the document belonging to each of the tested categories. Once retrieved the documents texts have been precessed by a pipeline in order to retrieve a normalised version of lexical chains. Afterwards, the lexical chain were subjected to the semantic expansion described in what follows. The normalization of textual representation of each Reuter documents, at a morphological and syntactic level, has been performed by the text-processing pipeline whose main phases based on specific tasks: *Sentence Segmentation* is responsible for breaking up documents (entity description, comments or abstract) into sentences (or sentence-like) objects which can be processed and annotated by ”downstream” components; *Tokenization* breaks sentences into sets of word-like objects which represent the smallest unit of linguistic meaning considered by a natural language processing system; *Lemmatisation* is the algorithmic process of determining the lemma for a given word. This phase substantially groups together the different inflected forms of a word so they can be analysed as a single item; *Stopwords elimination* phase filters out stop words from analysed text. Stop words usually refer to the most common words in a language, e.g. *the*, *is*, *at*, *which*, and so forth in English; *POS (Part-Of-Speech)-tagging* attaches a tag de-

noting the part-of-speech to each word in a sentence, e.g., *Noun*, *Verb*, *Adverb*, etc.; *Named Entity Recognition* phase categorizes phrases (referred to as entities) found in text with respect to a potentially large number of semantic categories, such as person, organization, or geopolitical location; *Coreference Resolution* phase identifies the linguistic expressions which make reference to the same entity or individual within a single document – or across a collection of documents. Once a normalized lexical chain for each documents has been obtained from the text-processing pipeline, its semantic expansion is built by exploiting the features of WordNet that will be described in detail in the following section.

4 WordNet-BASED DOCUMENT REPRESENTATION

The proposed approach uses WordNet in order to expand, at a semantic level, the lexical chains extracted from documents retrieved among the Reuters collection. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. In this context, we have defined and implemented a meta-model for WordNet to be exploited in the expanded lexical chains using a conceptualization as much as possible close to the way in which the concepts are organized and expressed in human language (Rinaldi, 2008). We consider concepts and words as graph nodes, whereas semantic, linguistic and semantic-linguistic relations as edges connecting nodes. For example, the *hyponymy* property is converted in an edge that links two concept nodes (nouns to nouns or verbs to verbs), while, a syntactic relation relates word nodes to word nodes. Concept and word nodes are considered with *DatatypeProperties*, which relate individuals with a predefined data type. Each word is related to the represented concept by the ObjectProperty *hasConcept* while a concept is related to words that represent it using the ObjectProperty *hasWord*. These are the only properties able to relate words with concepts and vice versa; all the other properties relate words to words and concepts to concepts. Concepts, words and properties are arranged in a class hierarchy, resulting from the syntactic category for concepts and words and from the semantic or lexical type for the properties. All elements have an ID within the WordNet offset number or a user defined ID. The semantic and lexical properties are arranged in a hierarchy. In Table 1 some of the considered properties and their domain and range of definition are shown.

Table 1: Properties.

Property	Domain	Range
hasWord	Concept	Word
hasConcept	Word	Concept
hypernym	NounsAnd VerbsConcept	NounsAnd VerbsConcept
holonym	NounConcept	NounConcept
entailment	VerbWord	VerbWord
similar	AdjectiveConcept	AdjectiveConcept

The use of domain and codomain reduces the property range application. For example, the hyponymy property is defined on the sets of nouns and verbs; if it is applied on the set of nouns, it has the set of nouns as range, otherwise, if it is applied to the set of verbs, it has the set of verbs as range. in Table 2 there are some of defined constraints and we specify on which classes they have been applied w.r.t. the considered properties; the table shows the matching range too.

Table 2: Model Constraints.

Costraint	Class	Property	Constraint range
AllValuesFrom	NounConcept	hyponym	NounConcept
AllValuesFrom	AdjectiveConcept	attribute	NounConcept
AllValuesFrom	NounWord	synonym	NounWord
AllValuesFrom	AdverbWord	synonym	AdverbWord
AllValuesFrom	VerbWord	also_see	VerbWord

Sometimes the existence of a property between two or more individuals entails the existence of other properties. For example, being the concept dog a hyponym of animal, we can assert that animal is a hypernymy of dog. We represent this characteristics by means of property features shown in Table 3.

Table 3: Property Features.

Property	Features
hasWord	*inverse* of hasConcept
hasConcept	*inverse* of hasWord
hyponym	*inverse* of hypernym; *transitivity*
hypernym	*inverse* of hyponym; *transitivity*
cause	*transitivity*
verbGroup	*symmetry* and *transitivity*

WordNet has been imported in Neo4J and afterwards visualized in Cytoscape (Shannon et al., 2003) according to a procedure similar to (Caldarola and Rinaldi, 2016) (Caldarola et al., 2016). Compared to the previous ones, this work focuses on the visualization of WordNet and the its most expensive part has consisted in defining a Cytoscape custom style to represent the *synonms rings* as tag clouds in an effective and clear way. We preferred to load WordNet objects from JWI APIs and serialize them in custom csv files to add some useful information in the csv lines, such

as the *word frequency* and the *polysemy* for the sake of the successive representation in Cytoscape. Before diving into the procedure details, it is worth clarifying the distinction between *synsets*, *synsets (or synonims) rings*, *index words* and *word senses*. As discussed in the previous section, a synset is a concept, i.e., an entity of the real world (both physical or abstract) whose meaning can be argued by reading the *gloss* definition provided by WordNet. Its meaning can be also understood by analysing the semantic relations linking it to other synsets or by reading the terms belonging to the synset (or synonims) ring. This one is a set of words (i.e. index words) generally used in a specific language (such as English) to refer that concept. The term synset itself is used to refer to set of synonyms meaning a specific concept. On the contrary, an index word is just a term, i.e., a *sign* without meaning; so, only when we link it to a specific concept we obtain a word sense, a word provided with meaning. An index word has got different meanings according to the context in which it is used and because of a general characteristic of languages: the *polysemy*. For example, the term *book* has eleven different meanings if it is used as noun (both lower and upper case), and so, it belongs to eleven different synsets. In addition to synsets glosses, WordNet gives some useful statistic information about the usage of the term *book* in each synset. The position of the term in each synonyms ring tell us how usual is the use of such term to mean that concept. The position of the term in each synset (comma separated in the listing) is a measure of the usage frequency of the term for each concept: higher the position, higher the frequency. Moreover, by counting the number of synsets which a term belongs to, it is possible to obtain its polysemy (e.g., the number of possible meanings of *book*). The information that we collect from WordNet for the synset nodes are the following: *Id*: the unique identifier for the synset; *SID*: the Synset ID as reported in the WordNet database; *POS*: the synset part of speech (POS); *Gloss*: the synset gloss which express its meaning; *Level*: the hierarchical level of synset in the whole WordNet hierarchy. While, for the word node we are interested in the following information: *Id*: the unique identifier for the word sense; *POS*: the word's part of speech (POS); *polysemy*: the word polysemy; *frequency*: the word frequency of the word sense as previously explicated. Finally, we retrieve the semantic links existing between synsets, by reporting the type of semantic link existing between them, e.g., *hyponym* or *meronym*, and the linguistic-semantic relations (*hasWord*), which connect word nodes to the corresponding synsets. Figure 2 shows the layout of the semantically-expanded lexical chains used in this work. Actually, we provide two layouts: the one is used for an high-level view of the expanded lexical chains and is used to get general insights from the document (and to decide about its main semantic category), while the other zooms in and provides details such as labels and IDs associated to nodes and edges. Focusing on the first layout (figure 2), it is possible to distinguish three types of nodes, depicted with three different colors: the white nodes represent words, the blue ones represent synset while the orange represent an original document word, i.e., a word that occurs in the Reuters document. Each word node is connected to the synset (synonyms ring) it belongs to a dashed line, this way making possible to visualize the synonym rings around the sysnet. In general, one synset may has one or more word nodes connected to it (due to the synonymy) and it is true also the contrary, i.e, one word nodes can be connected to one or more sysnets (due to the polysemy of such word). One sysnet can connect to other synsets through semantic relations (mostly *hyponym* but also *meronym*) depicted in dark green in the figure.

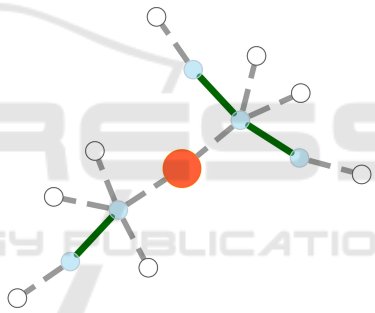

Figure 2: Reuters document structure.

5 GRAPH-BASED REPRESENTATION OF EXPANDED LEXICAL CHAINS

The defined meta-model to represent WordNet has been implemented in a labelled and properties-based graph within Neo4J and used in our context of interest. We applied the proposed model to construct the semantically-grounded expansions of a selection of one hundred documents taken from the Reuters-21578 corpus, which spans over ten different categories (or topics) such as: *earn*, *grain*, *trade*, *money*, *sugar*, *coffe*, *iron*, *cotton*, *meal* and *silver*. The considerations that have arisen and the discussions that follow, along with the images provided here, concern one of the Reuters corpus documents taken as an example, precisely, the number 6353. This latter belongs to the training set used in the modified Lewis

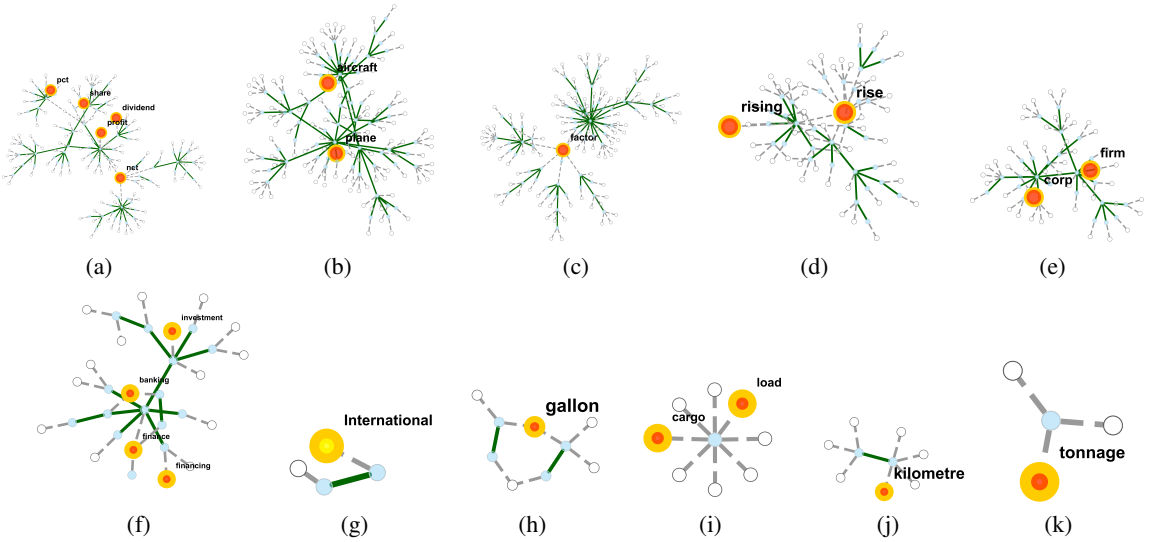

Figure 3: Reuters document graph sub components.

split (Lewis, 1992), which has been labelled as *earn*, which means earn on some commercial or business transaction, *earn* as net salary or wages, or as a profit or dividend in general. By using the proposed document representation strategy, we expect the user be able to recognize the document topic and its *specificity*, at a semantic level, just by having a look at the graph-based representation of the document and analysing its structure in terms of some topological characteristics like the *connectivity*, the number of *connected components* of the underlying graph, the number of the original terms (those coming from the Reuters document) belonging to each connected components, the spatial distribution of the original terms in each connected components. Figure 3 shows the graph sub components for the semantic expansion of Reuters no. 6353. Like almost all document expansions, it represents a disconnected graph to which several sub-components belong. For the sake of brevity and clarity, the figure depicts only eleven subcomponents, the remaining ones being too small to be worth dealing with. Each sub-component represents a connected arborescence of the graph that includes original terms that are close at a semantic level. In fact, the more the terms are close at a semantic level, the more likely they are close and connected at a topological level, due to the nature of the drawing algorithm (Force-directed graph drawing), which tries to reduce the crossing edges as much as possible and make edges of more or less equal length. Figures from 3(a) to 3(f) represent sub-components with a discrete dimension in term of nodes and edges, and refers to the main topics addressed by the document, i.e., *earn*, *aircraft*, *rise* or *rising*, *firm* and *bank*, while figure from 3(g) to 3(k) refer to marginal concepts like units of measure (*gallon*, *kilometre*, *tonnage*) and other related concepts like *cargo* and *load*. Our attention here focuses on the largest sub-component of the graph with the maximum number of original terms inside. This is the case for figure 3(a), which contains 70 synset nodes, 122 word nodes, 66 semantic relations (mostly hyponyms) and 127 meta-linguistic relations distributed over 192 nodes and 193 edges. Among the word nodes, there are five words contained in the original document, i.e., *pct*, which is the abbreviated form for *percentage*, *share*, *dividend*, *profit* and, finally, *net*. It worth to pointing out that the first four terms have a small degree of polysemy - the number of senses (meanings) the word can has according to the context - , w.r.t. *net*, which has 6 senses - for example *net* can mean *a computer network*, *the net income* (the case for this document) or *a trap made of netting to catch fish or birds or insects* and so fort. Accordingly, at a topological level, it results as a separated node w.r.t. the first. Moreover, while the first four nodes are mostly leaf node (with the exception of share, which has e greater polysemy) it represents a *bridge* nodes connecting three parts (semantically separated) of the connected sub-graph. Intuitively, terms with higher degree of polysemy do not allow us to recognize the topic or the *domain* of the document (even by reading the document we need to contextualize the text in order to attach the right sense to *net*), but, obviously, in this case, we attach the sense oriented to *North-West*, i.e, the meaning close to the most populated region of the sub-graph (that of *profit*, *dividend*, and so forth). The analysis of the other subcomponents of the semantic expansion, but one, leave no doubts about what is the main topic of the document. In fact, figures 3(b), 3(d) and 3(e) present sub-

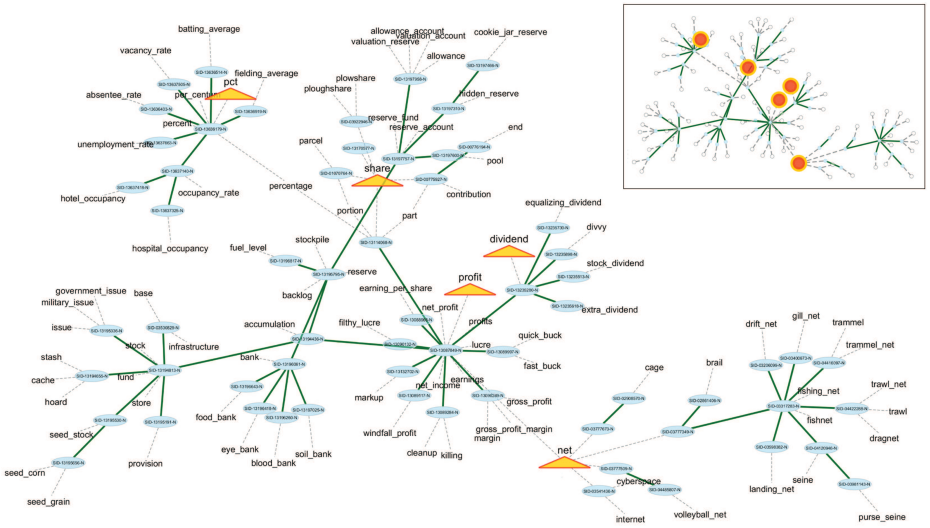

Figure 4: Reuter document semantic expansion sub-network.

graphs with a decreasing number of nodes and edges and, importantly, only two original terms (e.g., *aircraft* and *plane*). These sub-components are related to marginal concepts or topics addressed by the document, if we confront them w.r.t. sub-graph in figure 3(a). Furthermore, terms like *rise* or *rising* in figure 3(d) has an high degree of polysemy, 10 if considered as *Noun* and 17 if considered as *Verb*, so they do not help to recognize any topic by themselves. Even more so, figure 3(c) is useless in our analysis, because it contains only one original term, *factor*, which has a high degree of polysemy (7) and represent a kind of star-point between semantically separated regions. Figure 3(f) deserves some further considerations. It has 4 original terms belonging to *banking* and *finance* domains; furthermore, these terms are very close in the connected sub-graph, and also have a low degree of polysemy. They characterize the knowledge domain of the document as well as the terms in figure 3(a), but our strategy consider the *banking* topic as a second choice, due to the lesser number of original terms contained in corresponding sub-graph. This means that we can also define a ranking function between the sub-graph based on the number of original terms belonging to each sub-graph. Figure from 3(g) to 3(k) represent small connected components of the semantic expansion limited to one or at most two original terms. They give a small contribute to the identification of the document topic. Taking into account all the above considerations, it turns out that the sub-graph depicted in figure 3(a) represents that associated with the main topic of the document. Figure 4 shows the detailed representation of such sub-graph by putting in evidence the textual label associated to all words belonging to the synonyms set (synset). Each synset is represented with a blue oval while the original terms are depicted above orange diamonds. The figure clearly show the role of *bridge* for the word *net*, which has the greatest level of polysemy between the sub-graph words. Inside each oval is the sysnet ID retrieved from WordNet, furthermore, each synset is connected to the synonyms, represented as plain text, through a dashed line. The more remarkable observation that it worth to highlight to conclude this section is that the proposed strategy tries to recognize the document topic or domain by only representing the semantic-grounded expansion of the lexical chain underlying the document. All terms occurrences information like the term frequency (*tf*) or the inverse document frequency (*idf*) are neglected here in favour of a semantic and topological interpretation of the expanded lexical chain.

6 CONCLUSIONS

In this paper, a document representation methodology has been proposed and discussed at a qualitative level. We use a semantically-grounded graph models in order to visualize the more relevant terms in a document and the interconnections with semantically related terms. The implementation of our methodology within Neo4J results in a disconnected graphs containing several connected sub-graphs, each of them potentially referring to a topic or semantic category of the source document. The main addressed question is about the possibility to recognize the topics of a document just by analysing the topology of the graph underlying the expanded lexical chains by me-

ans of sub-graphs with the maximum number of original terms. The application of this methodology to approximately one hundred Reuters documents has demonstrated that if a predominant topic for the analysed document exists, a recurring pattern turns out, i.e., there exist a connected sub-graph with the maximum number of original terms extracted from the analysed document. Thus, it is possible to recognize the topic not in relation to the frequency of occurrence of terms, but in relation to topological characteristics of the graph, mainly the connectivity of the sub-graphs and their dimension. This strategy goes beyond the mere word-centric approach used in the most spread document representation model like the Space Vector Model because leaves aside the statistic of the document and suggests further researches in the topic detection field, which will be the subject of further studies.

REFERENCES

Begelman, G., Keller, P., and Smadja, F. (2006). Automated Tag Clustering: Improving search and exploration in the tag space. In *Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006*, Edinburgh, Scotland.

Caldarola, E., Picariello, A., and Rinaldi, A. (2016). Experiences in wordnet visualization with labeled graph databases. *Communications in Computer and Information Science*, 631:80–99.

Caldarola, E. G., Picariello, A., and Rinaldi, A. M. (2015). Big graph-based data visualization experiences: The wordnet case study. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, volume 1, pages 104–115. IEEE.

Caldarola, E. G. and 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*, pages 34–41. IEEE.

Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V. D. P., Loreto, V., Hotho, A., Grahl, M., and Stumme, G. (2007). Network properties of folksonomies. *AI Commun.*, 20(4):245–262.

Chen, Y.-X., Santamaría, R., Butz, A., and Therón, R. (2009). Tagclusters: Semantic aggregation of collaborative tags beyond tagclouds. In *Proceedings of the 10th International Symposium on Smart Graphics*, SG '09, pages 56–67, Berlin, Heidelberg. Springer-Verlag.

Fujimura, K., Fujimura, S., Matsubayashi, T., Yamada, T., and Okuda, H. (2008). Topigraphy: visualization for large-scale tag clouds. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1087–1088, New York, NY, USA. ACM.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611.

Hassan-Montero, Y. and Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. In *InScit2006: International Conference on Multidisciplinary Information Sciences and Technologies*.

Hu, X. and Wu, B. (2006). Automatic keyword extraction using linguistic features. In *Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, pages 19–23, Washington, DC, USA. IEEE Computer Society.

Kaptein, R. (2012). Using wordclouds to navigate and summarize twitter search. In *Proceedings of the 2nd European Workshop on Human-Computer Interaction and Information Retrieval*, pages 67–70. CEUR.

Lewis, D. D. (1992). *Representation and Learning in Information Retrieval*. PhD thesis, Computer Science Dept.; Univ. of Massachusetts; Amherst, MA 01003. Technical Report 91–93.

Lewis, D. D. (1997). Reuters-21578 text categorization test collection, distribution 1.0. *http://www. research. att. com/˜ lewis/reuters21578. html*.

Li, R., Bao, S., Yu, Y., Fei, B., and Su, Z. (2007). Towards effective browsing of large scale social annotations. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 943–952, New York, NY, USA. ACM.

McCandless, M., Hatcher, E., and Gospodnetic, O. (2010). *Lucene in Action: Covers Apache Lucene 3.0*. Manning Publications Co.

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

Rinaldi, A. M. (2008). A content-based approach for document representation and retrieval. In *Proceedings of the Eighth ACM Symposium on Document Engineering*, DocEng '08, pages 106–109, New York, NY, USA. ACM.

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. (2012). Improving tag clouds with ontologies and semantics. In *Database and Expert Systems Applications (DEXA), 2012 23rd International Workshop on*, pages 139–143. IEEE.

Rinaldi, A. M. (2013). Document summarization using semantic clouds. In *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pages 100–103. IEEE.

Rivadeneira, A. W., Gruen, D. M., Muller, M. J., and Millen, D. R. (2007). Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, pages 995–998, New York, NY, USA. ACM.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment

for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

Sinclair, J. and Cardew-Hall, M. (2008). The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–21.

Thorndyke, P. W. (1978). Knowledge transfer in learning from texts. In *Cognitive psychology and instruction*, pages 91–99. Springer.

Wei, Y. (2012). An iterative approach to keywords extraction. In *International Conference in Swarm Intelligence*, pages 93–99. Springer.

Yan, P. and Jin, W. (2012). Improving cross-document knowledge discovery using explicit semantic analysis. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 378–389. Springer.

Zhu, J., Wang, C., He, X., Bu, J., Chen, C., Shang, S., Qu, M., and Lu, G. (2009). Tag-oriented document summarization. In *Proceedings of the 18th international conference on World wide web*, pages 1195–1196, New York, NY, USA. ACM.