A Concept-Based Text Analysis Approach Using Knowledge Graph

Wenny Hojas-Mazo[1], Alfredo Simón-Cuevas[1], Manuel de la Iglesia Campos[1], Francisco P. Romero[2](✉), and José A. Olivas[2]

[1] Universidad Tecnológica de La Habana José Antonio Echeverría, Cujae, La Habana, Cuba
{Whojas, Asimon, miglesia}@ceis.cujae.edu.cu
[2] Universidad de Castilla-La Mancha, Ciudad Real, Spain
{FranciscoP.Romero, JoseAngel.olivas}@uclm.es

Abstract. The large amounts and growing of unstructured texts, available in Internet and other scenarios, are becoming a very valuable resource of information and knowledge. The present work describes a concept-based text analysis approach, based on the use of a knowledge graph for structuring the texts content and a query language for retrieving relevant information and obtaining knowledge from the knowledge graph automatically generated. In the querying process, a semantic analysis method is applied for searching and integrating the conceptual structures from the knowledge graph, which is supported by a disambiguation algorithm and WordNet. The applicability of the proposed approach was evaluated in the analysis of scientific articles from a Systematic Literature Review and the results were contrasted with the conclusions obtained by the authors of this review.

Keywords: Computational text analysis · Graph-based text representation Knowledge graph querying · Semantic processing

1 Introduction and Background

The large amounts and growing of unstructured texts, available in Internet and other information-centric application scenarios, are becoming a very valuable resource of information and knowledge. The effective processing and analysis of those text data sources to obtain the relevant information and knowledge has been an important challenge, due to this, the problem of text mining has gained increasing attention in recent years [2]. In order to address this challenge, the text data has been treated and processed using different representation levels, in most applications as bag-of-words or vector-space model, such as in information retrieval system. However, the information retrieval systems have traditionally focused more on facilitating information access rather than analyzing information to discover patterns and knowledge, which is the primary goal of text mining and the concept-based text analysis as specific task. In this sense, the graph-based text representation is emerging as a promising direction toward analyzing and exploiting the text structure [13], for example, the conceptual structure

J. Medina et al. (Eds.): IPMU 2018, CCIS 854, pp. 696–708, 2018.
https://doi.org/10.1007/978-3-319-91476-3_57

underlying. The use of graph-based text representation avoids the loss of structural and semantic information and reduced the text data scattering ratio. Once a text is represented as a graph, a variety of tools for graph analysis can be applied to perform quantitative and qualitive analysis of concepts, detecting closely contextually related concepts, identifying the key concepts that produce meaning, and perform integration of several text contents.

The useful of graph models in different text processing task, such as: topic labeling and detection [7, 12], text clustering [1], information retrieval [15, 18], text recommendation [5], and representation of linguistic information [24], have been reported. Several graph models applied to text representation, features and construction methods, have been reviewed in [7, 8, 11, 30]. However, the Concept Maps (CM) [23] is another graph model used to obtain the conceptual structure of a text [19, 25, 31], but very little exploited in the computational analysis of texts content. The CM is a graph-based knowledge representation, composed of concepts and labeled relationship between them that form propositions. The CM is a very useful and intuitive knowledge representation for capturing, representing and organizing the most significant of a topic and a set of conceptual meanings through of propositional structures [23]. Text analytics methods for extracting meaningful keywords and concepts facilitate content analysis, applying various technologies for capturing, processing, analyzing, and visualizing the immense volume, and variety of unstructured data from multiple textual sources [27]. Through CM, the large bodies of text are reduced to a relatively small number of concepts and relationship between them, so that a large corpus can be easily managed and understood via automatic concept mapping.

In this work, a Concept-Based Text Analysis Model (CTAM) is proposed. CTAM is based on the use of CM for structuring the texts content and a query language to retrieve relevant information and obtain knowledge from the constructed CM. CTAM is composed of two fundamental processes: *automatic concept mapping*, and *concept maps querying*. In the first process, a CM is automatically constructed from each text included in the texts collection to be analyzed, using a method based on the reported in [25]. The generated CM are stored in a *CM repository* (CMR). In the second process, an improved version of CMQL (Concept Maps Query Language) [28] is proposed for querying the CMR. CMQL provides the formalization of a set of different types of queries (*union*, *intersection*, *sub-map*, and *extension*) for exploring and mining a CMR from different perspectives, and offers more diversity of queries than the reported in [30]. In the search and integration tasks included in the query processing of CMQL, only syntactic aspects have been considered in the similarity analysis of the concepts. This constitutes a weakness due to the knowledge in CM is expressed in natural language, and several problems can be emerged due to the possible ambiguity of the concepts. For example, the not retrieval of useful and interesting information associated to concepts that are not syntactically similar to the included ones in the query, although they can be semantically similar, and the obtain of not appropriate results in the integration of associated information to semantically different concepts. To solve this weakness a semantic analysis method was included in this proposed approach, which is supported in a disambiguation algorithm and WordNet [20]. The word sense disambiguation in unstructured texts has been broadly studied, but there are few works approaching this problem in the CM context [6, 29]. The method reported in [29] improve the disambiguation results respect the reported in [6], through the use of heuristics based on *domain* (according to [4]), *context* and *gloss* and extending the context analysis process of the CM with other relations from WordNet. Nevertheless, the sequential application of these heuristics constitutes a limitation because the sense of the concept was determined according one of them and not taking advantage of the combination of the results obtained from each one. In this sense, a new disambiguation algorithm, based on [29], in which the results obtained for each heuristic are combined (inspired in [21]) to determine the more appropriate sense of the concept is also proposed.

The applicability of the CTAM was evaluated through the development of a case of study, in which 11 scientific articles from the Systematic Literature Review (SLR) reported in [9], were analyzed. The SLR is a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest [16]. The proposed CTAM offers a new approach of computational support to the analysis phase in the SLR context. In this case of study, different queries were carried out for analyzing the conceptual contents of those articles and to identifying relevant information that facilitated to obtain answer to the research questions outlined in that review. The results obtained were contrasted with the results and conclusions obtained by the authors of the reported review [9].

The rest of the paper is organized as follows: Sect. 2 describes the proposed Concept-Based Text Analysis Model and the defined processes; Sect. 3 presents the results of the developed case of study and the analysis carried out; and conclusions arrived and future works are given in Sect. 4.

2 Concept-Based Texts Analysis Model

CTAM is based on the use of the CM, which are automatically constructed from the texts included in a text collection and stored in a CMR, and the use of different types of queries defined in CMQL [28]. CMQL is applied to retrieve relevant information and obtain knowledge from the CMR. In this sense, two processes were defined: *automatic concept mapping*, and *concept maps querying*. Besides, a semantic analysis method to be applied for searching and integrating information in query processing from CMR, supported by a disambiguation algorithm and WordNet, was include in the last one process. An overview of the proposed approach is shown in Fig. 1.

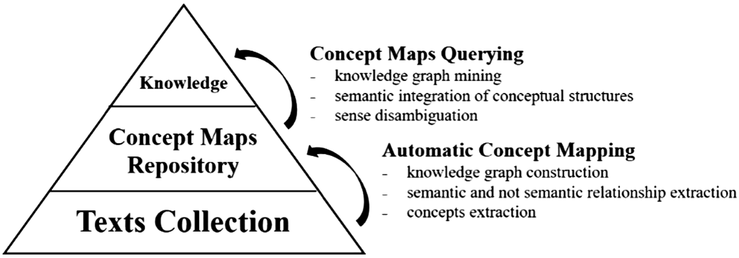
Knowledge
Concept Maps Repository
Texts Collection
Concept Maps Querying
- knowledge graph mining
- semantic integration of conceptual structures
- sense disambiguation
Automatic Concept Mapping
- knowledge graph construction
- semantic and not semantic relationship extraction
- concepts extraction

Fig. 1. Graphic overview of the proposed CTAM

2.1 Automatic Concept Mapping from Texts

In the proposed model, the automatic concept mapping is the process in which a CM is automatically constructed from each text included in the texts collection, which are stored in a CMR. This process is carried out through a method based on [25]. The method is conceived in three phases: *preprocessing*, *concepts extraction*, and *relationships extraction*. In the *preprocessing,* the text is segmented into sentences, and Freeling is used for obtaining the syntactic and grammatical information from the sentences. Next, several tasks are performed on each sentence, such as: tokens extraction, the morpho-syntactic and dependency analysis and the identification of named entities.

The *concepts extraction* phase is based on the identification of simple words or phrases (set of words), that by their composition can constitute a concept, through a set of lexical-syntactic patterns that have been defined for English and Spanish language [25]. The identification of concepts from external knowledge source, such as ontologies, is also included. A concept extracted list is obtained as result. The *relationships extraction* phase allows identifying explicit and implicit links between the previously extracted concepts using the information contained in the text, as well as that represented in an external knowledge source. The explicit relationships are extracted from each sentence using some lexical-syntactic patterns defined for this purpose [25]. The taxonomic relationships (implicit relationships) are extracted from BabelNet [22], and applying the *string matching technic* [14]. Other implicit relationships are extracted evaluating the proximity between two concepts in the text, according to [26]. The extraction of implicit relationships allows linking concepts that are not in the same sentence, which can be a useful information in contextual analysis task. Next, a refining process for eliminating redundancies or inconsistencies resulting from the application of lexical-syntactic patterns is performed. Finally, the concepts and propositions extracted are integrated in a CM. The combined use of the lexical-syntactic patterns for extracting the concepts and relationship between them, the identification of explicit and implicit links between the concepts and the use external knowledge sources allows to achieve a broad coverage of the textual content in the automatic construction of its conceptual representation (CM).

2.2 Concept Maps Querying Process

The CM mining process for retrieving relevant information and knowledge from the CMR is carried out through several types of queries defined in CMQL [28]. Through the different types of queries the system can retrieve information about concepts and propositions, with the results being shown by means of a CM or knowledge graph which is automatically constructed. The knowledge is produced when the captured concepts from different textual sources are integrated as part of the query results. The definition of the queries in CMQL is based on several mechanisms to filter and integrate concepts and propositional structures. This allows obtaining automatically different knowledge views from the CMR. In each query processing, the concepts and propositions including in the search source (set of selected CM) are processed as independent elements and, at the same time, they can be integrated through a semantic analysis process within the queries execution process (described below). Among the defined queries are: *union (CMUnion), intersection (CMInter),* and *projection (CMProj).*

The *union* query allows retrieve the knowledge graph that represents the concepts and the relationships between them extracted from a text collection, as a semantically integrated view of the concepts included in these different texts. The *intersection* query allows retrieve the knowledge graph that represents the common concepts included in the texts and the extracted relationships between them, which are represented in a certain amount of CM from the search source, according to the user interest. The minimum percent of CM in which a concept should be represented in the search source is defined as *support value* (*SV*). This numeric value is specified by the user when it executes the query. The *projection* queries allow retrieve the knowledge graph that represents the related concepts to a set of interest concepts (previously defined by the user) and the relationship between them from the search source of the query, considering different approaches and a maximum depth of R in the integrated graph of the search source. Three specifications of *projection* queries were included in this model to obtain different knowledge graph views: (1) considering only input link to the interest concept *c*; (2) considering only output link from *c*; and (3) combining both types of links. The first two types of queries are very useful to analyze the authority or centrality levels of *c* with respect to other related concepts. This is a new approach to the application of Kleinberg's concepts [17] in the conceptual analysis of textual contents. In the case of R = 0 (when *SV* is not specified by the user), it is assumed that the interest of the user is to identify if the any interesting concepts are included in the search source and if there are any relationship between them.

A refinement of the mathematical formalization of the queries in CMQL is presented in Table 2. This formalization allows to obtain an abstract model, independent of the implementation language of the queries and the storage format of the CM. Before describing the queries, let us consider the following symbolism in Table 1.

Table 1. Symbolism for mathematical formalization of the queries

Symbols	Definitions
SS	Set of *CM* $\{CM_1, CM_2, \ldots, CM_n\}$ defined as the search source of a query
CM^q	Concept map obtained as a result of a query $q/q = \{U, I, Proj\}$
IC	Set of *interest concepts* defined by the user (needed in projection query)
P	Set of propositions $p = (c_{o;}\ c_{d;}\ lp)$
c_o	Origen concept in a proposition *p*
c_d	Destiny concept in a proposition *p*
lp	Linked phrase used for labeled the relationship between two concepts
R	Path length between two concepts in a CM
$INC_{CM}^{R}(c)$	Set of concepts included in all the paths of length *R* from the concept *c* in a CM, considering input link to *c* (relative to the authority level of *c*)
$OUTC_{CM}^{R}(c)$	Set of concepts included in all paths of length *R* from the concept *c* in a CM, considering output link from *c* (relative to the centrality or hub level of *c*).

Table 2. Mathematical formalization of Concept Maps Query Language

Queries	Mathematical formalization
CMUnion	$CMUnion(SS) = (\cup\ CM_i \mid CM_i \in SS), = (C^U, P^U)$ where $C^U = \mid n > 1$ and $P^U = \mid n > 1.$
CMInter	$CMInter^{SV}(SS) = (\cap\ CM_i \mid CM_i \in SS), CM^I = (C^I, P^I)$ where $C^I = \mid n > 1$ and $P^I = \mid (n > 1, p_j \in P^I, and\ c_{o,}\ c_d \in C^I)$
CMProj	$CMProj^R(SS, IC) \subseteq CMUnion(SS)$ and is defined as:
	1. $CMProj^{R,\ IN}(SS, IC) = CM^{Proj} = (C^{Proj}, P^{Proj})$ where $C^{Proj} \mid (c \in IC, CM^U = CMUnion(SS))$, and $P^P = \mid(\mathrm{p_j} = (c_{o;}\ c_{d;}\ lp)\ and\ c_o, c_d \in C^{Proj}$
	2. $CMProj^{R,\ OUT}(SS, IC) = (C^{Proj}, P^{Proj})$ where $C^{Proj} \mid (c \in IC, CM^U = CMUnion(SS))$, and $P^{Proj} = \mid (\mathrm{p_j} = (c_{o;}\ c_{d;}\ lp)\ and\ c_o, c_d \in C^{Proj}$
	3. $CMProj^R(SS, IC) = CM^{Proj} = CMProj^{R,\ IN}(SS, IC) \cup CMProj^{R,\ OUT}(SS, IC)$

In CMQL [28], the information is recovered through identifying syntactic equivalency between the concepts included in the query and the contents in the search source, and the associated semantics to these concepts is not considered. The same thing happens in the process of integration that is carried out as part of the queries processing. Nevertheless, the concepts and propositions in the CMR can be subjected to ambiguity in many cases, because they are expressed in natural language and the ambiguity is an inherent characteristic of the language. Therefore, the effectiveness in the CM querying process can be limited if a semantic analysis task is not included the queries processing; being the identification of the most rational sense of the concepts an important aspect. In the following section, the semantic analysis method that has been proposed in the CTAM is described.

2.3 Semantic Analysis in the Querying Process

The proposed semantic analysis method is based on a process of semantic extension of concepts, and a set of rules in the integration, search and retrieval tasks included in the CMR querying. In this method, the semantic information associated to the concepts is captured from WordNet and a concept sense disambiguation algorithm is used for reducing the ambiguity that may emerged. The semantic extension is applied to all concepts included in a CMR and is defined as the process of associating to one concept other synonym terms identified in WordNet. Initially, the *synsets* in which each concept appears in WordNet are recovered, and then are classified in: *ambiguous* - *AC* - (those having more than one associated *synsets*), *not ambiguous* - *NA* - (only one associated *synset*) or *unknown* - *UC* - (not associated *synset*). Next, a disambiguation algorithm is applied to identify the most appropriated sense (or senses) for the ambiguous concepts. This algorithm, based on [29], improves the disambiguation results, fundamentally through combining the results obtained for each heuristic for determining the sense of the concept. This method is inspired in [21]. After applying the disambiguation algorithm, the lists of *ambiguous* and *not ambiguous* concepts are updated and each one of those concepts are extended with the terms included in their associated *synset*. The algorithm is defined as follows (Table 3):

Table 3. Concept sense disambiguation algorithm

Input: an ambiguous concept (c_a); CM (in which c_a is included); the set of *synsets* of c_a ($S(c_a)$).
Output: the more appropriated *synsets* for disambiguating c_a.

1. Preprocessing: For each concept $c_i | c_i \neq c_a$ and linking phrase (e) of CM
   a. the set of *synsets* of c_i ($S(c_i)$) and e ($S(e)$) are obtained from WordNet.
   b. the n most representative domains associated to the *synsets* included in ($S(c_i)$) and ($S(e)$) are identified and stored in the set D_{mc}, according to their occurrence frequency in those *synsets* and considering $n \leq 5$;
2. Domain analysis: For each $s_i | s_i \in S(c_a)$, the influence degree ($h_d(s_i)$) that each domain $d \in D_{cm}$ exercises on the sense s_i is calculated, through the sum of the occurrence frequencies of each d associated to $s_{i;}$
3. Context analysis: For each $s_i | s_i \in S(c_a)$, the influence degree ($h_c(s_i)$) that context (propositional structure in which c_a appear) exercises on the sense s_i is calculated as follows:
$$h_c(s) = w_c * \frac{\sum_{c_i \in C_r} \mathrm{rel}(s, c_i)}{|C_r|} + w_r * \frac{\sum_{r_i \in R_r} \mathrm{rel}(s, r_i)}{|R_r|}$$
where C_r and R_r are the set of concepts and linking phrases, respectively, included in a vicinity of radius r having c_a as it center, and *rel(s, e)* is a value indicating the semantic relatedness between the *synset s* and each *synset* associated to the concepts and linking-phrases included in the context. In the formula, w_c and w_r are weights assigned according to the desired relevance degree that information from concepts and linking phrases will have, respectively, for the heuristic. The sum of w_c and w_r should always be 1 to guarantee a maximum possible value of $h_c(s)$ of 1.
4. Gloss analysis: For each $s_i | s_i \in S(c_a)$, the influence degrade ($h_g(s_i)$) of definition (gloss) of the *synset* s_i in WordNet, considering its relation with the elements of context of c_a, is calculated as follows:
$$h_g(s) = w_c * \frac{|C_r \cap G(s)|}{|C_r|} + w_r * \frac{|R_r \cap G(s)|}{|R_r|}$$
being G(s) the set of words from the gloss of synset s. The weights w_c and w_r have the same purpose described in the previous step.
5. Heuristics combination: For each $s_i | s_i \in S(c_a)$, the global influence of the different heuristics ($h_{dcg}(s_i)$) is calculated as follows:
$$h_{dcg}(s_i) = w_d h_d(s_i) + w_c h_c(s_i) + w_g h_g(s_i)$$
where w_d, w_c and w_g are weights with values representing the influence degree that each heuristic has in the precision of the full algorithm, which were defined according to precision results obtained by domain, context and gloss heuristics reported in [29]
6. Selection of the resulting *synset*. The more appropriated *synset* for c_a is the *synset* having a higher $h_{dcg}(s_i)$. In case of more than one *synset* having condition, all of them are considered, and the other ones are discarded.

The semantic integration task is aimed at explicitly integrating propositional structures (initially disconnected) through the unification of concepts (in a unique node) represented in different CM and it is applied when the query is performed on more than one CM. The concepts unification process is carried out through the identification of synonymous concepts in the selected CM as the search source of the query, and using several rules (R). Considering that $S(c_i)$ is the set of *synsets s* associated to a concept c_i and c_1 and c_2 are two concepts included in different CM, then c_1 and c_2 are unified if:

- *R1:* $(c_1, c_2 \in NAC) \wedge (S(c_1) = S(c_2))$*; or*
- *R2:* $(c_1, c_2 \in AC) \wedge (\exists s' | s' \in S(c_1) \wedge s' \in S(c_2))$*; or*
- *R3:* $((c_1 \in NAC \wedge c_2 \in AC) \vee (c_1 \in AC \wedge c_2 \in NAC)) \wedge (\exists s' | s' \in S(c_1) \wedge s' \in S(c_2))$*; or*
- *R4:* $(c_1, c_2 \in UC) \wedge (c_1 = c_2)$*; (fundamentally included for unifying not included concepts in WordNet, for example named entities)*

As result, if *R4* was triggered or the labels of c_1 and c_2 (in the case of other triggered rules) the same label of these concepts is used for representing the *unified concept* in the query results. In other cases, the label used for representing the unified concept is constructed with the labels of both concepts separated by a comma and enclosed in [ ] (ex. [c_1, c_2]). Finally, the *synset* associated to the *unified concept* is decided according to: (1) *if R1 was triggered, then the synset is the same to the* c_1 or c_2; (2) *if R2 was triggered, then the synsets are the common ones between the associated to* c_1 and c_2; (3) *if R3 was triggered, then the synset is the one associated to the* $c_i \in NAC$.

The semantic analysis is also considered in the proposed retrieval model, specifically in the projection queries (*Q*). In the process of query specification, the definition of one or more interesting concepts (*CQ*) by the user is required, besides defining the search source (*SS*) selecting a set of CM from the CMR. The selected CM are integrated through a union query as internal task in the query processing. Therefore, SS can be formally defined by the tuple (C^{ss}, P^{ss}), where C^{ss} is the set of concepts and P^{ss} is the set of propositions, included in the selected CM. Several rules were defined for identifying if a concept $c_j / c_j \in C^{ss}$ is retrieved or not, from a concept $c_i \in CQ$, where syntactic and semantic analysis are combined. These rules are described below and are executed following the same order in which they appear. However, it is possible to parameterize the combination of the analysis type considered, according to: using the syntactic analysis, using the semantic analysis, or combining both analyses. Being a concept $a/a \in CQ$, a concept $b/b \in C^{ss}$, $T(c_i)$ the set of words included in the label of c_i, and $ST(c_i)$ the set of synonym terms included in $S(c_i)$. The concept b is retrieved from SS if: (R1) $a \equiv b$ *(syntactically equivalent);* or (R2) $a \in T(b)$*;* or (R3) $a \in ST(b)$*.*

3 Applicability of CTAM: Case of Study

The evaluation of the proposed model turns out complex because a method for this purpose has not been identified. Nevertheless, in this section we present the case of study carried out in order to show the applicability of CTAM in the SLR context [16]. Much of the SLR processes requires high time-consuming, and several manual tasks [3], implying a great effort when mixing evidence from multiple studies and synthesizing evidence across studies, fundamentally in the analysis phase. On the other hand, some barriers (most of them affects the analysis phase) have been identified for carried out this type of review, such as [3]: lack of support for data extraction and analysis, difficulties of summarizing and aggregating data (especially qualitative data), difficulties of mixing evidence from multiple studies, difficulties for synthesizing evidence across studies, among others. Precisely, the proposed approach contributes to reduce the time consumption and the negative effects of some of those barriers. In order to demonstrate the applicability of our approach, a real SLR reported in [9] was selected. In this review, 11 scientific articles from 1820 primary studies were selected as the most relevant evidences to be analyzed.

The case of study was carried out using a collection with 11 texts, which were constructed using the abstract, introduction and conclusions from those articles analyzed in [9], in a similar way as the reported in [10]. Those texts have an average of 822 words and 33 sentences. Two different queries from CMQL were executed to analyze the content included in those texts and to identify relevant information that facilitated to give answer to the research questions reported in [9]. Specifically, the *CMInter* and *CMProj* were selected for supporting this analysis phase. Through of *CMInter* queries, relevant concepts and relationship between them from the texts can be retrieved, mixing evidence from these multiple studies and synthesizing the contents. In this sense, several *SV* for exploring the evidences, such as: 80% (Q1), 70% (Q2), 60% (Q3) and 50% (Q4), were used in the *CMInter* queries. Through of *CMProj* queries, relevant and useful information for answering the research questions can be retrieved, selecting some identified keywords in these questions as interest concepts.

According the definition of CTAM, initially the concept mapping process was carried out and a CM was automatically constructed from each text. Next, four *CMInter* queries with the different *SV* mentioned were executed on the constructed CMR. The domain specific concepts (DSC) and the generic contextual terms (GCT) identified as relevant keywords in [9] were used to measure the precision, recall and F-measure of retrieved concepts. The results are shown in Table 4, in which the measures were evaluated for the keywords sets: DSC, GCT and DSC + GCT. In Fig. 2, the resultant CM of the *CMInter* (SV = 50%) query is shown. The size of concepts represents the frequency in CMR, therefore [software system, software] and [dependability, reliability] are the most relevant retrieved concepts. The *strong relation* indicates that the concepts [framework, model] and 'ISO' are contextually related, suggesting further analysis by the reviewer. In this example, the results of the proposed semantic analysis method are also illustrated, through the integration of some syntactically different concepts, for example: 'framework'-'model' and 'dependability'-'reliability', because a synonymy relationship was automatically identified among them. This can help reviewers to quickly know the terminologies used in the articles to refer at the same concept.

In addition, the *CMProj*[1] query was executed using the constructed CMR as search source and '*standard*' as the interest concept. The result is shown in Fig. 3. The

Table 4. Results in the keywords identification task using *CMInter* queries

	Precision			Recall			F-measure		
	DSC	GCT	DSC + GCT	DSC	GCT	DSC + GCT	DSC	GCT	DSC + GCT
Q1	100	0	100	37.5	0	16.7	54.5	0	28.6
Q2	80	20	100	37.5	10	22.2	51.1	13.3	36.4
Q3	57.1	28.6	85.7	37.5	20	27.8	45.3	23.5	42
Q4	62.5	25	87.5	37.5	20	27.8	46.9	22.2	42.2
Ave.	**74.9**	**18.4**	**93.3**	**37.5**	**12.5**	**16.8**	**49.5**	**14.8**	**37.3**

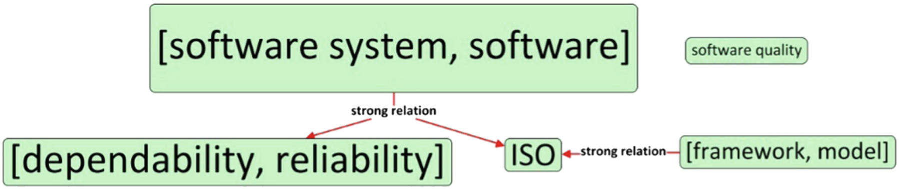

Fig. 2. Result of the *CMInter* query (SV = 50%)

objective of this query is to obtain useful information for answering one of the research questions answered in [9]: "*Which software reliability models have been developed by following the recommendations in International Standards?"*. In this question, '*standard*' is one of the most relevant terms on which is necessary to retrieve information.

Through the selected query it is possible to retrieve those strongly related concepts with the term '*standard*', including concepts associated to '*reliability models*' and '*International Standards*'. The Fig. 3 shows several retrieved concepts that they represent different international standards, such as: *ISO*, *ECSS*, *IEEE*, *SQuaRE*, *COSMIC* and *Space Standardization*, most of them (66,6%) were also identified in the manual analysis carried out by Febrero et al. [9]; although all of them are represented in Fig. 3. The analysis of the evidences to answer the research question can be enriched applying others *CMProj* queries, for example, increasing the R value and using others interest concepts, such as: '*reliability*'. As the results of this case of study, several beneficial aspects of the application of CTAM to the exploration, interpretation, and decision-making in the analysis phase of primary studies in a SLR were emerged, such as: obtaining the relevant concepts from the articles that should be reviewed; quickly knowing the terminologies associated to the concepts include in the different articles; assisting reviewers to know which concepts are contextually related to keywords from the research questions; facilitating the qualitative data (concepts and relationship between them) mining and its analysis; and mixing evidences from multiple studies.

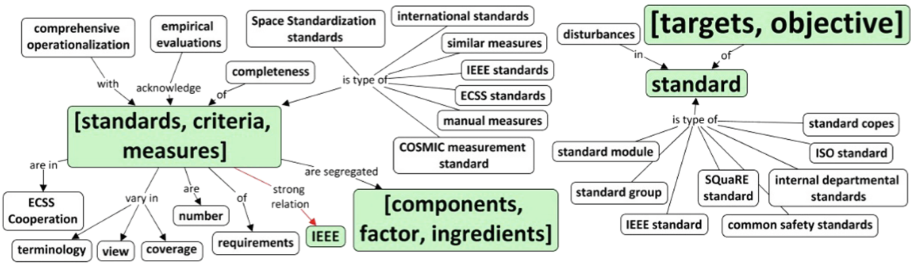

Fig. 3. Result of the *CMProj*[1] query using *standard* as interest concept

4 Conclusions and Future Works

This paper presented a new concept-based text analysis approach, based on the use of CM to represent the conceptual structure underlying of the texts content and an improvement version of CMQL, to retrieve relevant information and obtain knowledge from the conceptual structure represented. On the other hand, the resulting CM from each query provides in CMQL constitutes conceptual and summarized representation views of the content included in the texts. The integration of the proposed semantic analysis method, supported in WordNet and the use of a disambiguation algorithm, to the query processing defined in CMQL allowed improve the results of search and information integration. The results of the case of study carried out to evaluate the applicability of the proposed approach demonstrated several benefits to the conceptual exploration, interpretation, and decision-making in the review of primary studies carried out in a SLR. In future works, others graph operations will be considered to extend the proposed model and increasing the results of the concept-based texts analysis through the automatic detection of frequent patters and topics from the CMR.

Acknowledgments. This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant MERINET: TIN2016-76843-C4-2-R (AEI/FEDER, UE)

References

1. Abdulsahib, A.K., Kamaruddin, S.S.: Graph based text representation for document clustering. J. Theor. Appl. Inf. Technol. **76**(1), 1–10 (2015)
2. Aggarwal, C.C., Zhai, C.X. (eds.): Mining Text Data. Springer, New York (2012). https://doi.org/10.1007/978-1-4614-3223-4
3. Al-Zubidy, A., Carver, J.C., Hale, D.P., Hassler, E.E.: Vision for SLR tooling infrastructure: prioritizing value-added requirements. Inf. Softw. Technol. **91**, 72–81 (2017). https://doi.org/10.1016/j.infsof.2017.06.007
4. Bentivogli, L., Forner, P., Magnini, B.; Pianta, E.: Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In: Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources, pp. 101–108 (2004). https://doi.org/10.3115/1706238.1706254
5. Benkoussas, C., Bellot, P.: Information retrieval and graph analysis approaches for book recommendation. Sci. World J. **2015**, 1–8 (2015). https://doi.org/10.1155/2015/926418
6. Cañas, A.J., Leake, D.B., Maguitman. A.G.: combining concept mapping with CBR: towards experience-based support for knowledge modeling. In: Proceedings of FLAIRS Conference, pp. 286–290. AAAI Press (2001)
7. Chang, J.Y., Kim, I.M.: Research trends on graph-based text mining. Int. J. Softw. Eng. Appl. **8**(4), 147–156 (2014). https://doi.org/10.14257/ijseia.2014.8.4.16
8. Chen, L., Jose, J.M., Yu, H., Yuan, F.: A semantic graph-based approach for mining common topics from multiple asynchronous text streams. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1201–1209 (2017). https://doi.org/10.1145/3038912.3052630

9. Febrero, F., Calero, C., Moraga, M.A.: Software reliability modeling based on ISO/IEC SQuaRE. Inf. Softw. Technol. **70**, 18–29 (2016). https://doi.org/10.1016/j.infsof.2015.09.006
10. Felizardo, B.K.R., Andery, G.F., Paulovich, F.V., Minghim, R., Maldonado, J.C.: A visual analysis approach to validate the selection review of primary studies in systematic reviews. Inf. Softw. Technol. **54**, 1079–1091 (2012). https://doi.org/10.1016/j.infsof.2012.04.003
11. Hassan, G.S., Abdulsahib, A.K., Kamaruddin, S.S.: Graph-based text representation: a survey of current approaches. Res. J. Appl. Sci. Eng. Technol. **14**(9), 334–340 (2017). https://doi.org/10.19026/rjaset.14.5073
12. Hulpus, I., Hayes, C., Karnstedt, M., Greene, D.: Unsupervised Graph-based Topic Labelling Using Dbpedia. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 465–474, ACM (2013). https://doi.org/10.1145/2433396.2433454
13. Indurkhya, N.: Emerging directions in predictive text mining. WIREs Data Min. Knowl. Disc. **5**, 155–164 (2015). https://doi.org/10.1002/widm.1154
14. Jiang, X., Tan, A.H.: CRCTOL: a semantic-based domain ontology learning system. J. Am. Soc. Inform. Sci. Technol. **61**(1), 150–168 (2010). https://doi.org/10.1002/asi.21231
15. Karim, G., Mouna, T.K., Lynda, T., Maher, B.J.: Graph-based methods for significant concept selection. Procedia Comput. Sci. **60**, 488–497 (2015). https://doi.org/10.1016/j.procs.2015.08.170
16. Kitchenham, B.A., Charters S.: Guidelines for performing systematic literature reviews in software engineering. Technical report EBSE 2007-001, Keele University and Durham University Joint Report (2007)
17. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM (JACM) **46**(5), 604–632 (1999). https://doi.org/10.1145/324133.324140
18. Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., Lawley, M.: Graph-based concept weighting for medical information retrieval. In: Proceedings of the 17th Australasian Document Computing Symposium, pp. 80–87 (2012). https://doi.org/10.1145/2407085.2407096
19. Kowata, J.H., Cury, D., Silva, M.C.: Concept maps core elements candidates recognition from text. In: Proceedings of the 4th International Conference on Concept Mapping, 1, pp. 120–127 (2010)
20. Miller, G., Fellbaum, C. (eds.): WordNet: An Electronic Lexical Database. The MIT Press, Cambridge (1998)
21. Navigli, R.: Word sense disambiguation: a survey. ACM Comput. Surv. **41**(2), 1–69 (2009). https://doi.org/10.1145/1459352.1459355
22. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012). https://doi.org/10.1016/j.artint.2012.07.001
23. Novak, J.D., Cañas, A.J.: The theory underlying concept maps and how to construct them. Technical report IHMC CmapTools 2006-01, 32502, USA (2006)
24. Pinto, D., Gómez, H., Vilariño, D., Singh, V.K.: A graph-based multi-level linguistic representation for document understanding. Pattern Recogn. Lett. **41**, 93–102 (2014). https://doi.org/10.1016/j.patrec.2013.12.004
25. Rodríguez, A., Simón, A.: Método para la extracción de información estructurada desde textos. Revista Cubana de Ciencias Inf. **7**(1), 55–67 (2013)
26. Rodríguez, A., Simón, A., Guevara, E., Hojas, W.: Modelo de representación de textos basado en grafo para la minería de texto. Ciencias de la Inf. **46**(1), 63–71 (2015)

27. Sasson, E., Ravid, G., Pliskin, N.: Creation of knowledge-added concept maps: time augmention via pairwise temporal analysis. J. Knowl. Manage. **21**(1), 132–155 (2017). https://doi.org/10.1108/JKM-07-2016-0279
28. Simón, A., Ceccaroni, L., Rosete, A., Suárez, A., Victoria, R.: A support to formalize a conceptualization from a concept maps repository. In: Proceedings of the 3rd International Conference on Concept Mapping, pp. 68–75 (2008)
29. Simón, A., Ceccaroni, L., Rosete, A., Suárez, A., de la Iglesia, M.: A concept sense disambiguation algorithm for concept maps. In: Proceedings of the 3rd International Conference on Concept Mapping, pp. 14–21 (2008)
30. Sonawane, S.S., Kulkarni, P.A.: Graph based representation and analysis of text document: a survey of techniques. Int. J. Comput. Appl. **96**(19), 1–8 (2014). https://doi.org/10.5120/16899-6972
31. Valerio, A., Leake, D., Cañas, A.J.: Using automatically generated concept maps for document understanding: A human subjects experiment. In: Proceedings of 5th International Conference on Concept Mapping, pp. 438–445 (2012)