Context-informed Knowledge Extraction from Document Collections to Support User Navigation

Mario Cataldi[1], Claudio Schifanella[1], K. Selçuk Candan[2]
Maria Luisa Sapino[1], Luigi Di Caro[1]
[1]Università degli Studi di Torino
10149 Torino
Italy
{cataldi,schi,mlsapino,dicaro}@di.unito.it

[2]Arizona State University
Tempe, AZ 85283
USA
candan@asu.edu

ABSTRACT: *Most of the existing document and web search engines rely on keyword-based queries. To find matches, these queries are processed using retrieval algorithms that rely on word frequencies, topic recentness, document authority, and (in some cases) available ontologies. In this paper, we propose an innovative approach to exploring text collections using a novel keywords-by-concepts (KbC) graph, which supports navigation using domain-specific concepts as well as keywords that are characterizing the text corpus. The KbC graph is a weighted graph, created by tightly integrating keywords extracted from documents and concepts obtained from domain taxonomies. Documents in the corpus are associated to the nodes of the graph based on evidence supporting contextual relevance; thus, the KbC graph supports contextually informed access to these documents. The construction of the KbC graph relies on a spreading-activation like technique which mimics the way the brain links and constructs knowledge. In this paper, we also present CoSeNa (Context-based Search and Navigation) system that leverages the KbC model as the basis for document exploration as well as contextually-informed media integration.*

Keywords: Knowledge Management, HCI, Navigation System, Keywords Proximity

Received: 2 March 2010, Revised 8 April 2010, Accepted 11 April 2010

1. Introduction

Popular methods for text retrieval are mostly based on available feature statistics (Salton and Buckley, [23]). Some recent systems also leverage available semantics to guide the retrieval process towards an equilibrium between relatedness and wisdom (Gauch et al., [191). In this work, we propose CoSeNa, an innovative system to help the users navigate within text collections, relying on a novel keywords-by- concepts (KbC) graph. The KbC of a text collection, in the context of a given domain knowledge, is a weighted graph constructed by integrating the domain knowledge (formalized in terms of domain taxonomies, i.e., the semantic context) with the given corpus of text documents (i.e., the content). Consequently, unlike related works, where the feature weights either reflect the keyword statistics in the database or the structural relationships between the concepts in the taxonomies, the weights in the KbC graph reflect both the semantic context (imposed by the taxonomies) and the documents' content (imposed by the available document corpus[1]).

[1]*In the news application that motivates this research, this corpus is defined by the temporal frame of interest and/or the keywords appearing in the news articles.*

Figure 1 shows a fragment of a sample KbC graph. This example integrates geographical domain knowledge (extracted from a taxonomy, which organizes geographic entities of the World - cities, provinces, regions, states, continents) and the keywords extracted from a collection of newspapers articles. In this example, the newspaper articles from which the keywords are extracted are about the "*9/11 World Trade Center terrorist attack*" and the "*American invasion of Afghanistan*":

• Each node in the graph is either a concept from the domain taxonomy, or a keyword extracted from the content of the document base.

• The graph is bipartite: each edge connects a domain concept to a content keyword (hence the name *keywords-by-concepts* graph). The edges are weighted and they weigh the strength of the relationship between the connected nodes in the given context. In Figure 1, the weights of the edges are visually represented through the thickness of the edges.

Consider the geographical concepts "*US*" and "*Afghanistan*". In the graph fragment, "*US*" is linked to the content keywords "*terrorism*", "*Bin Laden*", "*U.N.*", and "*nuclear*" (in decreasing order of weights[2]), while "*Afghanistan*" is connected to "*terrorism*" and "*Bin Laden*". Thus, these last two keywords create a content-based association between the two geographical concepts "*US*" and "*Afghanistan*", which in the original domain taxonomy would appear far from each other (they belong to different continents). In CoSeNa, while browsing in the document space, the user can leverage such associations as bridges between concepts.

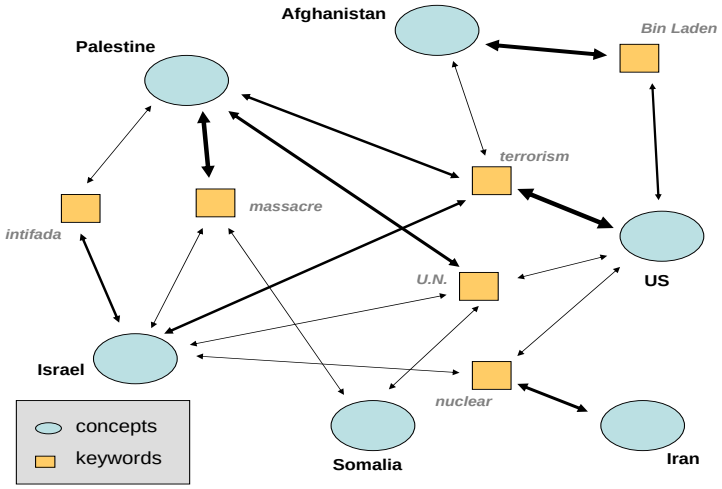

Figure 1. An example KbC graph constructed using concepts from a geographical domain taxonomy and keywords extracted from a corpus of news documents

Figure 2. A sample screenshot of the Navigation interface of CoSeNa engine

[2]*We will discuss in Section 3.5 how these content keywords are extracted and how the links are defined*

Using CoSeNa (Figure 2), the user can navigate within the document space by starting from any concept or keyword. At each step, CoSeNa presents the user navigational alternatives as well as documents that are relevant in the given context. Navigational possibilities are represented relying on the tag cloud metaphor: the font sizes express the strength of the relationships among concepts and keywords. Documents associated to the concepts and keywords are enumerated in a ranked list. When the user clicks on a document, the system shows the corresponding document and highlights the contextually important concepts and keywords in the document. The user can navigate into the KbC space by clicking on the concepts and the keywords highlighted in the tag clouds as well as in the documents. The proposed system also provides integration with three on-line popular media sources: Google Maps, Flickr, and YouTube. To achieve context-based integration, CoSeNa queries the content sources leveraging the concepts and keywords in the clouds and presents the results to the user in a unified interface.

1.1 Contributions of this Paper

In this paper, we recognize that the assumption that users know what they want precisely is not always valid. Also, the conventional way of presenting the user a list of candidate documents may fail to help the user observe the contextual relationships, among the concepts and documents, hidden in the database. Therefore, traditional feedback processes, which can be degraded significantly if the user feedback is uninformed or inconsistent, may fail to be effective.

This problem can be addressed to a limited extent by relying on domain taxonomies that can inform the user about the domain specific relationships among concepts and, thus, support relatively more informed navigation within the document space (Sacco, [22]). However, most taxonomies describe the given domain with categories and relationships which are valid at the time at which the taxonomy was created. In our work, we note that (especially in dynamically evolving domains, such as newspapers) document contents themselves are very real-world context-aware, since they in fact reflect what people know and are interested in.

For example, let us reconsider the concepts "US" and "Afghanistan" in Figure 1. Given the shape of the corresponding nodes, we can see that they are concepts from a given taxonomy (which we know, in this case, is the input geographical taxonomy). Importantly, this taxonomical domain knowledge does not change over time. Yet, before the 9/11 events, very few people would immediately associate "Afghanistan" and "US". After the 9/11 events, however, keywords, such as "terrorism" and "Bin Laden" would strongly link "US" and "Afghanistan". Thus, domain-specific taxonomies, when used alone, cannot be effective in capturing and leveraging the evolving semantics associated to the concepts. In particular, *keywords associated to the same concept would strongly differ at different times* because the background contexts about the places, people, and the facts are different. Taxonomies alone cannot capture this.

Thus, in this paper, we propose to address these deficiencies of traditional purely feedback-based and purely taxonomy-based solutions, by implementing an innovative exploration and navigation approach which discovers and highlights hidden, contextually-relevant relationships between concepts as well as keywords characterizing documents in the corpus. More specifically,

- we propose a novel *keywords-by-concepts* (KbC) graph, which is a weighted graph constructed by a tight integration of available domain taxonomies (i.e., the semantic context) with the keywords extracted from the documents' search space (i.e., the content);

- we assign the weights of the edges in the KbC graph to reflect both the keyword statistics in the database as well as the semantics and structural relationships between the concepts in the taxonomies relying on a spreading activation like technique;

- we present a novel *concept-expansion* strategy leveraging the document context, imposed by the available document corpus, in disambiguating the semantic context described by the input taxonomies; and

- we leverage the KbC graph in the CoSeNa (*Context-based Search and Navigation*) system for context-aware navigation and document retrieval. We show that the KbC graph can not only be used for locating the most relevant documents within a given context, but can also be used for (a) explaining why two concepts in the given domain are related and (b) for identifying dominant concepts and keywords in the given context.

The rest of the paper is organized as follows: Section 3 reports the algorithm used to define the semantic correlations among concepts and keywords and explains how to build a KbC navigational graph. Section 4 defines how to bind the most relevant documents to each graph node, using the semantic information inferred by the graph. Section 5 shows the features of the implemented system.

2. Related Work

The problem of indexing text collections is becoming more important than ever with the explosion of web contents. Most current information retrieval (IR) systems rely on a keyword search scheme, where queries are answered relying on the keyword contents of the text, sometimes also relying on available taxonomies.

In this section we present an overview of the existing approaches to manage this task.

2.1 Query Expansion

A major challenge with IR is that user queries are often under-specified: users tend to provide at most 2-3 keywords and this is often insufficient to hone on the most relevant documents (Spink et al., [26]). In the web, where hyperlinks provide structural evidence to help identify authoritative sources, link analysis is used to help tackle this problem. Even then, however, query under-specification remains a significant challenge. Query expansion is one of the most popular approaches to address this challenge. Mandala et al. [15] present an overview of the common techniques. Generally speaking, the goal is to modify the initial query by adding, removing and changing terms with similar ones.

Existing state-of-the-art query expansion approaches can mainly be classified into two classes: global and local analysis. The first one relies on corpus wide statistics such as the co-occurrences of every pair of terms. To expand a query, the terms which are more similar to it are also considered. On the other hand, the local analysis uses only some initially retrieved documents for further query expansion. An hybrid solution was introduced by Xu and Croft [28], where both co-occurrences and local concepts selection are used for retrieval.

Sparck Jones [25] presents an example of global analysis approach: the authors proposed a technique which adds terms obtained from term clusters built based on co-occurrences of terms in the document collection. A similar work was presented by Qiu and Frei [19]; in this work term similarities are used instead of term co-occurrences. These methods, however, can not handle ambiguous terms: if a query term has different meanings, the clustering will add non-related terms thus making the expanded query ambiguous as well. Qiu and Frei [19] introduce a method for expanding target concepts of the whole query instead of a term-by-term change.

Cui et al. [4] present a way of exploiting user logs for query expansion. Every search engine website accumulates a large amount of query logs. The basic idea is that if a set of documents is usually selected for the same queries, then the terms in these documents are somehow related to the terms of the queries. Thus some correlations between query terms and document terms can be established based on such query logs.

2.2 Dimensionality Reduction

Another approach relies on dimensionality reduction techniques like Latent Semantic Indexing (LSI) (Deerwester et al.,[5]); Furnas et al., [10]). LSI decomposes terms and documents in a shared low dimensional space which captures the latent information contained in the corpus. This is obtained through the matrix decomposition Singular Value Decomposition (SVD) (Eckart and Young, [6]). Although this method is known to be more effective than standard frequency-based query systems, it suffers of the same problem of terms ambiguity.

2.3 Spreading Activation

The concept of spreading activation represents an iterative process that builds semantic networks based on the propagation (or spread) of information between the nodes. While this theory finds several applications in cognitive science, artificial intelligence, psychology and biology, it has also been used for information retrieval (Crestani, [3]). The first study in this direction was done by Preece [17], who he associated the model to IR. Recent approaches use this method for improving keyword search results by connecting query terms with supported phrases, synonyms, and documents (Aswath et al., [1]).

2.4 Retrieval by Navigation

An alternative approach to retrieval is to rely on an exploratory process instead of document indexing and query matching (Li and Candan, [14]). As stated by Shashank [24] there exist three reasons for preferring this retrieval by navigation approach over pure keyword-based text retrieval:

- Query formulation represents the most critical step in the whole retrieval process (Rieh and Xie, [20]) because of a variety of factors like user inexperience and lack of familiarity with terminology;
- In many cases the scope of a user query is too broad to express precisely using a set of keywords.
- Sometimes users prefer to navigate within a topic rather than being despatched to some system-relevant documents. Navigation process helps users understand the surrounding context to better hone on the relevant documents. This behavior is called orienteering (Teevan et al., [27]).

Consequently, even if many existing retrieval systems continue to rely on the more traditional query-based IR model, there is a recent tendency towards relying on browsing in contrast to directed searching. In these schemes, querying is nothing but an initial way of identifying starting points for navigation, and navigation is guided based on the context supplied in the query as well as any additional semantic metadata, such as taxonomies. These "semi-directed or semi-structured searching" processes (Ellis, [7]; Shashank, [24]) help address the "don't-know-what-I-want" behavior (Bates, [21]) more effectively than relevance feedback schemes that assume that the user knows what she wants.

2.5 Relevance Feedback

A well known query reformulation method is *user relevance feedback* (Ruthven and Lalmas, [2]). The idea is to ask the users to mark relevant documents in search results and re-weighting the keywords of the initial query based on how effective they are according to such feedback. The obvious drawback of this technique is that it puts significant overhead on the users and assumes that the users know what they want and can provide consistent feedback. Since this is rarely the case, the relevance feedback may by ineffective or may require significant amount of interactions.

One way to reduce the load on the user is to rely on *pseudo relevance* feedback (Grootjen and van der Weide, [12]) (also known as blind feedback), where the top ranked documents are assumed to be relevant and query enrichment is performed without user intervention, using these top-ranked documents. This scheme, however, works only if the initial query results are indeed highly relevant and can degenerate if the first query results contain not-so-relevant documents. Query reformulation can also be done by replacing items in a thesaurus with their longer descriptions. The thesaurus may be based on the used collection or based on a top domain knowledge like Wordnet (Fellbaum, [8]). However, these schemes still assume that user's initial query is highly precise and its expansion is sufficient to identify the relevant documents.

3. Construction of Keywords-by-Concepts (KbC) Graph

In this section, we describe how to create a *keywords-by-concepts* (KbC) navigational graph to support the exploration of the data, by highlighting the keyword and concept relationships, given a domain taxonomy H and a corpus of documents (contents) D. The construction algorithm combines information coming from a structural analysis of the relationships formalized in H with the analysis of the most frequent keywords appearing in the corpus of documents D. In the resulting graph, the weighted edges connecting keywords and concepts provide context-based navigation opportunities. In Section 5, we will show the use of the graph in assisting navigation and exploration within CoSeNa system.

The construction of the graph is preceded by a 4-steps analysis process, which extracts, from the given taxonomy and document corpus, the information needed to identify the concept-keyword mappings relevant in the given context:

1. The first step maps the concepts in the input taxonomy onto a concept-vector space in a way that encodes the *structural* relationships among nodes in the input taxonomy. The embedding from the concept hierarchy to the concept-vector space is achieved through a concept propagation scheme which relies on the semantical relationships between concepts implied by the structure of the taxonomy to annotate each concept node in the hierarchy with a concept-vector (Section 3.1).

2. The second step pre-processes the corpus of documents, to extract keyword frequencies (Section 3.2).

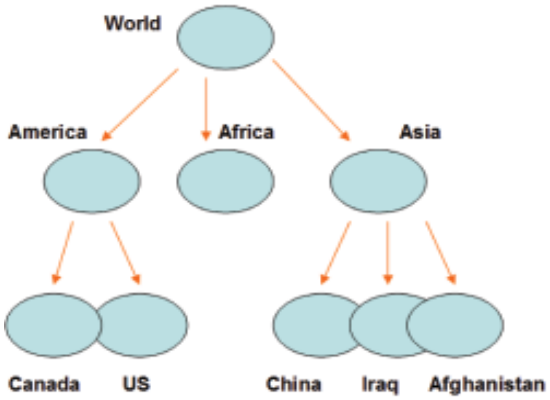

Figure 3. A geographical taxomomy fragment

3. For each document in the considered data set, the third step identifies the set of concepts that best describe that document. This process is based on the similarities between concept-vectors and document-vectors (Section 3.3).

4. For each concept, the last step extracts the most relevant keywords contained in the documents classified under it (Section 3.4). This helps identify highly correlated concepts and keywords, providing the basis for the keywords-by-concepts (KbC) navigational graph construction.

At the end of these four steps, it is possible to construct the KbC graph (Section 3.5) by leveraging the semantic similarities among keywords (from the corpus) and concepts (from the taxonomy).

Next, we discuss these steps in detail.

3.1 Taxonomy Vectorization: Embedding Concepts into a Concept-Vector Space

In order to support discovery of mappings between concepts and documents (which are represented as keyword-vectors), we also map concepts in the given domain taxonomy, H (*C*,*E*), onto a concept-vector space. More specifically, given a taxonomy, H (*C*,*E*), with $n = |C|$ concepts, we represent each concept node as a vector *cv* with *n* dimensions such that each vector represents the semantical relationship of the corresponding concept node with the rest of the nodes in the taxonomy. For this analysis step, we rely on the CP/CV mapping process proposed by Kim and Candan (2006). Given a taxonomy, CP/CV assigns a *concept-vector* to each concept node in the taxonomy, such that the vector encodes the *structural* relationship between this node and all the other nodes in the hierarchy. The spreading activation like process works as follows:

• each concept-vector of the nodes is simply initialized by setting to 1 the weight corresponding to itself; i.e., considering the node c_i in the given hierarchy, the initial concept-vector of this node is

$$\overrightarrow{cv_{c_i}} = \langle\ 0,...,1,...0\ \rangle \qquad (1)$$

where the only non-zero weight is associated with the i-th dimension related to che node c_i. The total number of dimensions is equal to the number of the nodes in H(*C*,*E*).

• Then, the process repeatedly enriches the concept-vectors of the nodes by enabling neighboring nodes to exchange concept weights. The propagation of the weights works by adding to each concept-vector the weights of the neighbour ones (parent and children), multiplied by a propagation degree that sets how much information has to migrate from one node to the neighbours. The *propagation degree* is computed in a way that reflects the local structure of the taxonomy (Kim and Candan, 2006).

This process is iterated until all nodes are informed of all the others. Consider, for example, the taxonomy fragment (containing nine concept nodes) presented in Figure 3. CP/CV maps each concept into a 9-dimensional vector (Table 1). Vectors' elements are associated to the taxonomy nodes, considered in breadth first order. In particular, for example, the root is represented by the vector

$$\langle 0.45,\ 0.169,\ 0.141,\ 0.158,\ 0.018,\ 0.018,\ 0.018,\ 0.021,\ 0.021 \rangle \qquad (2)$$

	world	Asia	Africa	America	Afghanistan	Iraq	China	Canada	US
$\overrightarrow{cv}_{world}$	0.450	0.169	0.141	0.158	0.018	0.018	0.018	0.021	0.021
$\overrightarrow{cv}_{Asia}$	0.052	0.469	0.006	0.006	0.156	0.156	0.156	0.0003	0.0003
$\overrightarrow{cv}_{Africa}$	0.100	0.012	0 .873	0.012	0.0006	0.0006	0.0006	0.0007	0.0007
$\overrightarrow{cv}_{America}$	0.057	0.007	0.007	0.520	0.0003	0.0003	0.0003	0.204	0.204
$\overrightarrow{cv}_{Afghanistan}$	0.004	0.100	0.0002	0.0002	0.872	0.012	0.012	0	0
$\overrightarrow{cv}_{Iraq}$	0.004	0.100	0.0002	0.0002	0.012	0.872	0.012	0	0
$\overrightarrow{cv}_{China}$	0.004	0.100	0.0002	0.0002	0.012	0.012	0.872	0	0
$\overrightarrow{cv}_{Canada}$	0.006	0.0003	0.0003	0.165	0	0	0	0.806	0.023
$\overrightarrow{cv}_{US}$	0.006	0.0003	0.0003	0.165	0	0	0	0.023	0.806

Table 1. The concept-vectors, related to the taxonomy in Figure 3, calculated by applying the algorithm proposed by Kim and Candan (2006).

in which the first component (the one associated to the tag "*world*"), dominates over the others that contribute to the definition of the concepts. The second, third and fourth components reflect the weight of "*Asia*", "*Africa*" and "*America*" respectively in the semantic characterization of "*world*", while the remaining components represent the weights of the three descendants of "*Asia*" and of the two descendants of "*America*".

$\overrightarrow{lv}_{world}$	{government, country, president, administration, leader · · ·}
$\overrightarrow{lv}_{Asia}$	{technology, crisis, sale, economy, area · · ·}
$\overrightarrow{lv}_{Africa}$	{UN, disease, World Bank, peacekeeper, tutsi, · · ·}
$\overrightarrow{lv}_{America}$	{democrat, economy, war, agreement, nominee · · ·}
$\overrightarrow{lv}_{Afghanistan}$	{taliban, Osama Bin Laden, terrorism, embassy, government, · · ·}
$\overrightarrow{lv}_{Iraq}$	{Saddam Hussein, weapon, missile, persian gulf, UN, · · ·}
$\overrightarrow{lv}_{China}$	{World Trade Organization, administration, Bill Clinton, Chen Shui Bian, Jiang Zemin, · · ·}
$\overrightarrow{lv}_{Canada}$	{traveler, Quebec, airfare, Ontario, mountain · · ·}
$\overrightarrow{lv}_{US}$	{George Bush, president, Al Gore, congress, Bill Clinton· · ·}

Table 2. The evidence-vectors ($\overrightarrow{lv}$) obtained using the articles extracted from the New York Time data set (described in Section 5). The presented terms are ordered based on the corresponding weights which are omitted in the figure for clarity.

3.2 Text Document Vectorization: Mapping Documents into Keyword-Vectors

In this step, given a data corpus *D* of documents (also defined contents), we analyze and extract a representative document-vector for each of them.

Thus, each of the $m = |D|$ documents are represented with a *document-vector* in which each component represents a keyword. As usual, the keyword extraction includes a preliminary phase of stop-word elimination and stemming. For the stemming process, we used Wordnet (Fellbaum, [8]).

The weight of each keyword is computed using augmented normalized term frequency (Salton and Buckley, [23]). Unlike the standard term frequency (TF), this gives credit to any term that appears in the corpus and adds some additional credit to terms that appear more frequently. In this way we preserve the keywords that appear less frequently. The reason for this is

that too often the most relevant terms related to an argument (or category, in our case) are specific keywords that do not appear so frequenly in a big corpus of document: the augmented normalized term frequency permits us to preserve this information.

In short, given a corpus document d_i, we define the related document-vector as

$$\vec{d_i} = \{w_{i,1}, w_{i,2}, \ldots, w_{i,v}\} \quad (3)$$

where v is the size of the considered vocabulary, and $w_{i,j}$ is the normalized term frequency of the j - *th* vocabulary term in the i - *th* document.

3.3 Analysis of Concepts Describing a Given Document

The concept-vectors assigned to the concept nodes provide a convenient way to identify best concepts describing each of the given documents:

• Input to this step are
  – the set $CV = \{\vec{cv}_1, \ldots, \vec{cv}_n\}$ of the concept-vectors representing the taxonomy and
  – the set $DV = \{\vec{dv}_1, \ldots, \vec{dv}_m\}$ of vectors representing the documents to be described in terms of concepts in the taxonomy.

Keyword-vectors of the documents are defined in the space of the entire set of document keywords; each dimension corresponds to a keyword, and the weights in the vector represent the relevance of the corresponding keyword value in the document represented by the vector.

• Output of this analysis step are *sets of representative concepts* associated to the documents in the corpus. We capture this notion of representativeness through the similarity among the taxonomy and document-vectors representing taxonomy concepts and documents, respectively.

Semantic similarities (at the basis of the association process) between the concepts and the documents being associated are computed by

• unifying the vector spaces of the concepts and the vector space of the documents. The unification of the spaces consists in unioning dimensions in the given ones, and representing every vector in the new extended space by setting to 0 the values corresponding to those dimensions that were not appearing in the original vector space, while keeping all the other components unchanged.

• using the cosine similarity of the vectors.

For each document in the corpus, the concepts that best describe the document are those concepts whose similarities with the document are above an adaptively computed critical point. The steps of this discovery process are as follows: for each document $d_j \in D$ we

1. consider the document-vector $\vec{dv}_j$
2. compute its similarity wrt. all the concept-vectors describing the given taxonomy.

$$sim(\vec{sv}_i, \vec{dv}_i) = \sum_{i=1}^{u} \vec{cv}_i[k] \cdot \vec{dv}_j[k] \quad (4)$$

3. sort the concept-vectors in decreasing order of similarity wrt. $\vec{dv}_j$;
4. choose the cut-off point to identify the concepts which can be considered *sufficiently similar*. Our method adaptively computes this cut-off as follows:

It

(a) first ranks the concepts in descending order of match to $d\vec{v}_j$, as previously calculated;

(b) computes the *maximum drop* in match and identifies the corresponding drop point;

(c) computes the *average drop* (between consecutive entities) for all those nodes that are ranked before the identified maximum drop point;

(d) the first drop which is higher than the computed *average drop* is called the critical drop. We return concepts ranked better than the point of critical drop as candidate matches.

At the end of this phase, each document in D has a non-empty set of concepts associated to it.

3.4 Discovery of Concept-Keyword Semantic Relationships

The next step towards the KbC construction process is to discover the concept-keyword mappings using these associations identified in the previous step. In other words, in this phase, we find those keywords that relate strongly to the concepts in the taxonomy.

Let cv_{c_i} denote the concept-vector corresponding to concept c_i. We denote the set of documents described by the concept c_i as $D_{desc}(\vec{cv}_{c_i})$. Notice that, in general, the sets of associated documents for different concepts are not disjoint, since the same document can be assigned to multiple (similar) concept-vectors. Note also that, at the end of the process, some of the concept nodes of the taxonomy may notbe associated as a descriptive concept to any of the documents in the database. For such concepts, the corresponding sets, D_{desc}, of associated documents are empty.

At this step, given a concept *ci* and the set, $D_{desc}(\vec{cv}_{c_i})$, of associated documents, we search for the most contextually informative keywords corresponding to this concept. More specifically, we compute the degree of matching between the given concept and a keyword which occurs in the associated documents by treating

- the set of documents in $D_{desc}(\vec{cv}_{c_i})$ which contain the keyword as positive evidence of relationship between the concept and the keyword within the given context, and
- the documents in the database containing the keyword but not associated tothe concept as negative evidence against the relationship.

Intuitively, this is analogous to treating (a) the concept-vector corresponding to the concept c_i as a query and (b) the set of associated documents as positive relevance feedback on the results of such query. Recognizing this, given a concept c_i and a corresponding set of associated documents, $D_{desc}(\vec{cv}_{c_i})$, we identify the weight, $u_{i,j}$, of the keyword k relying on a probabilistic feedback mechanism (Ruthven and Lalmas, 2003):

$$u_{i,j} = \log \left[\frac{\frac{r_{i,j}}{(R_i - r_{i,j})}}{\frac{(n_j - r_{i,j})}{N - n_i - R_i + r_{i,j}}} \right] \times \left| \frac{r_{i,j}}{R_i} - \frac{n_j - r_{i,j}}{N - R_i} \right| \qquad (5)$$

where:

- $r_{i,j}$ is the number of documents in $D_{desc}(\vec{cv}_{c_i})$ containing the keyword k_j;
- n_j is the number of documents in the corpus containing the keyword k_j;
- R_i is the number of documents in $D_{desc}(\vec{cv}_{c_i})$; and
- N is the number of documents in the corpus.

The first term increases as the number of the associated documents containing the keyword k_j increases, while the second term decreases when the number of the non-associated documents containing the keyword k_j increases. Therefore, keywords that are highly common in a specific association and not much present in others will get higher weights.

For each concept, we consider all keywords contained in at least one document.

We apply an adaptive cutoff to this set in order to select those keywords with the highest weights. Given concept c_i, the selected keywords and their weights are collected in a so-called *evidence-vector,* $\overrightarrow{lv_{c_i}}$.

Table 2 shows the most relevant terms included in the evidence-vectors related to the taxonomy fragment presented in Figure 3.

3.5 Constructing the KbC Graph Leveraging the Concept-Keyword Mappings

At the end of the previous phases, for each concept c_i, we have obtained an evidence-vector,

$$\overrightarrow{lv_{c_i}} = \langle u_{i,1}, u_{i,2}, \ldots, \rangle \tag{6}$$

$\overrightarrow{clv_{world}}$	{government, country, US, president, America · · ·}
$\overrightarrow{clv_{Asia}}$	{technology, crisis, China, sale, economy · · ·}
$\overrightarrow{clv_{Africa}}$	{UN, disease, World Bank, peacekeeper, tutsi, · · ·}
$\overrightarrow{clv_{America}}$	{ US, democrat, economy, war, agreement · · ·}
$\overrightarrow{clv_{Afghanistan}}$	{taliban, US, Osama Bin Laden, terrorism, Iraq, · · ·}
$\overrightarrow{clv_{Iraq}}$	{Saddam Hussein, weapon, missile, US, persian gulf, · · ·}
$\overrightarrow{clv_{China}}$	{World Trade Organization, administration, US, Bill Clinton, Chen Shui Bian, · · ·}
$\overrightarrow{clv_{Canada}}$	{traveler, Quebec, airfare, Ontario, mountain · · ·}
$\overrightarrow{clv_{US}}$	{George Bush, president, Al Gore, congress, Bill Clinton· · ·}

Table 3. The combined-vectors ($\overrightarrow{clv}$), related to the taxonomy presented in Figure 3, obtained using the previously created evidence-vectors ($\overrightarrow{lv}$) and concept-vectors ($\overrightarrow{cv}$). The presented terms are ordered based on the corresponding weights (again, omitted in the figure for clarity). Terms that are in bold are picked from the considered taxonomy.

that encodes the related keywords in the corpus and their weights. In this final phase of the KbC construction, we link together the concepts and keywords using these relationships.

Let $C = \{c_1, \ldots, c_n\}$ be the set of concepts in the input taxonomy, H, and $K = \{k_1, \ldots, k_m\}$ be the set of all keywords appearing in it at least one *evidence-vector*. We construct KbC, in the form of an undirected, node-labeled, edge-weighted graph, $G(V_C \cup V_K, E, l, \rho)$, as follows:

- Let V_C be a set of vertices, $V_C = \{v_{c_i}, \ldots, v_{c_n}\}$, where vertex $v_{c_i} \in V_C$ is las labeled as "k_j"; i.e., $l(v_{k_j}) =$"k_j"; and
- For all $v_{c_i} \in V_C$ and $v_{k_j} \in V_K$ such that $\overrightarrow{lv_{c_i}}[j] = \emptyset$, there exists an edge $\langle v_{c_i}, v_{k_j} \rangle \in E$ such that

$$\rho(\langle v_{c_i}, v_{k_j} \rangle) = \rho_{i,j} = \frac{\overrightarrow{lv_{c_i}}[j]}{\| \overrightarrow{lv_{c_i}} \|} \tag{7}$$

Therefore $\rho_{i,j}$ represents the relative weight of the keyword k_j in the corresponding vector $\overrightarrow{lv_{c_i}}$, i.e. the role of the keyword k_j in the context defined by the concept v_{c_i}.

4. Unification of the Concept and Keyword Vector Spaces

In order to support exploration of the documents in the corpus CoSeNa needs to associate, for each node of the KbC graph, a corresponding (ranked) list of documents. In Section 3.3, we have already described how to associate descriptive concepts to the documents. This initial mapping between concepts and documents, however, relied only on the semantic context provided by the taxonomy (captured by the concept-vectors, $\vec{cv}$), but did not account for the context implied by the document corpus (captured by the collection of evidence-vectors, $\vec{lv}$). Thus, before we obtain the final mapping between concepts and the documents, we need to enrich the concept-vectors, which represent the structured knowledge, with the help of the evidence-vectors, which represent the real-world background knowledge.

4.1 Associating Combined-Vectors to the Concepts in the given Taxonomy

At this point, for each concept c_i, we have two vectors: (a) the concept-vector, $\vec{cv}_{c_i}$, representing the concept-concept relationships in the corresponding taxonomy and (b) the evidence-vector, $\vec{lv}_{c_i}$, consisting of keywords that are significant in the current context defined by the corpus. In order to combine the concept and the collection evidence-vectors, into a single combined-vector,

$$\vec{clv}_{c_i} = \alpha_{c_i}.\vec{cv}_{c_i} + \beta_{c_i}.\vec{lv}_{c_i} \quad (8)$$

we need to first establish the relative impacts (i.e. α_{c_i} and β_{c_i}) of the taxonomical knowledge versus real-world background knowledge.

As defined in Section 3.3, let D_{desc} $\vec{cv}_{c_i}$) be the set of documents for which the concept c_i is a good descriptive concept. Also, given concept,c_i, let

- $S(\vec{cv}_{c_i})$ be the set of documents resulting from querying the database using the concept-vector, $\vec{cv}_{c_i}$; and
- $S(\vec{lv}_{c_i})$ be the set of documents obtained by querying the database using the evidence-vector, $\vec{lv}_{c_i}$.

We quantify the relative impacts, c_i and c_i, of the concept and evidence-vectors, $\vec{cv}_{c_i}$ and $\vec{lv}_{c_i}$, by comparing how well $S(\vec{cv}_{c_i})$ and $S(\vec{lv}_{c_i})$ approximate $D_{desc}(\vec{cv}_{c_i})$.
In other words, if

- $C_{c_i} = D_{desc}(\vec{cv}_{c_i}) \cap S(\vec{cv}_{c_i})$ and
- $L_{c_i} = D_{desc}(\vec{cv}_{c_i}) \cap S(\vec{lv}_{c_i})$,

then we expect that

$$\frac{\| \alpha_{c_i}.\vec{cv}_{c_i} \|}{\| \beta_{c_i}.\ \vec{lv}_{c_i} \|} = \frac{|C_{c_i}|}{|L_{c_i}|} \quad (9)$$

If the concept and extension-vectors are normalized to 1, then we can rewrite this as

$$\frac{\alpha_{c_i}}{\beta_{c_i}} = \frac{|C_{c_i}|}{|L_{c_i}|} \quad (10)$$

Also, if we further constrain that the combined-vector clv_{c_i} is also normalized to 1,

$$\| \alpha_{c_i}.\ \vec{cv}_{c_i} + \beta_{c_i}.\ \vec{lv}_{c_i} \| = 1, \quad (11)$$

then, solving these equations for α_{c_i} and β_{c_i}, we obtain:

$$\alpha_{c_i} = \frac{|C_{c_i}|}{|C_{c_i} + L_{c_i}|} \quad \beta_{C_i} = \frac{|L_{c_i}|}{|C_{c_i} + L_{c_i}|} \tag{12}$$

Thus, given a concept, c_i, we can compute the corresponding combined-vector as

$$\overrightarrow{clv_{c_i}} = \frac{|C_{c_i}|}{|C_{c_i} + L_{c_i}|} \cdot \overrightarrow{cv_{c_i}} + \frac{|L_{c_i}|}{|C_{c_i} + L_{c_i}|} \tag{13}$$

Table 3 shows an example that reports the combined-vectors related to the taxonomy fragment presented in Figure 3; as described in this section, the evidence-vectors are therefore integrated by using the structural information provided by the previously created concept-vectors.

4.2 Associating Combined-Vectors to the Keywords in the given Corpus

In order to associate combined-vectors to the keywords extracted from the given corpus, we consider the keywords concept neighbors in the corresponding KbC graph. By construction, each keyword node $v_{k_j} \in V_k$ in the KbC graph is connected to at least one concept node, $v_{c_i} \in V_C$. Thus, the combined-vector for $\overrightarrow{clv_{k_j}}$ is computed as

$$clv_{c_i} = \sum_{c_i \in neighbor(v_{k_j})} \frac{\rho_{i,j}}{\| \overrightarrow{clv_{c_i}} \|} \cdot \overrightarrow{clv_{c_i}} \tag{14}$$

where $\rho_{i,j}$ is the strength of the relationship between concept c_i and keyword k_j obtained through taxonomy and corpus analysis in Section 3.5. As it is the case for the $\overrightarrow{clv_{c_i}}$ vectors, $\overrightarrow{clv_{k_j}}$ are also normalized to 1.

4.3 Associating Documents to KbC Nodes in the given Context

Since, at this point, each concept and keyword node in the KbC graph has its own combined-vector $\overrightarrow{clv}$, the documents in the given corpus can be associated under these nodes as in Section 3.3, but using $\overrightarrow{clv}$ vectors instead of $\overrightarrow{cv}$ vectors. In this manner, using the combined-vectors, CoSeNa is able to associate to each concept and keyword, not only the documents that contain that concept or the keyword, but also the documents containing all contextually relevant concepts and keywords.

4.4 Measuring Concept-Concept and Keyword-Keyword Similarities in the given Context

At this point, each concept and keyword node in the KbC graph has an associated combined-vector $\overrightarrow{clv}$, capturing both the taxonomical relationships between concepts and the context defined by the documents in the given corpus. Therefore, inaddition to associating documents to KbC nodes, the similarities between concept and keywords in the given context (defined by the taxonomy and the document corpus) can be measured using the cosine similarities between these vectors. In fact, since all the concepts are mapped into the same vector space, the knowledge expressed by each node is comparable with all the others; i.e, it is possible to compute semantic similarities by simply using the cosine similarity measure.

In the next section, we will describe use of this in CoSeNa to support document exploration.

5. CoSeNa System Overview

In this section, we present an overview of the CoSeNa system, which leverages the KbC model introduced in this paper. With CoSeNa system, the user can navigate through the nodes in the KbC graph (computed in a preliminary pre-processing phase), starting from any concept or keyword. At each step, CoSeNa presents the user navigational alternatives as well as documents that are relevant in the given context. Navigational alternatives are represented relying on the tag cloud metaphor: given a concept or keyword,

• the system identifies most related concepts and keywords (using the KbC graph and concept-concept/keyword-keyword similarities), and

• forms a concept cloud (consisting of related concepts) and a keyword cloud (consisting of related keywords).

Concept and keyword font sizes express the strength of the relationships among concepts and keywords. Documents associated to the concepts and keywords are enumerated in a list ordered with respect to the weights calculated in Section 4.3. When the user clicks on a document, the system shows the corresponding document and highlights the contextually important concepts and keywords in the document. The user can navigate into the KbC space by clicking on the conceptsand keywords highlighted in the tag clouds as well as in the documents.

Figure 4. Navigation search initiated with the geographical concept "*Iraq*"

CoSeNa also provides media integration with GoogleMaps, Flickr, and YouTube: in order to retrieve related materials, CoSeNa queries the content sources using the semantic associations reported in the term clouds.

5.1 Search and Navigational Interface

Figures 4 and 5 show the use of the CoSeNa system in a scenario, where a corpus of news documents (the New York Times articles collection, which contains 300,000 text entries with over 100,000 unique keywords[3]) is explored with the help of a geographical concept taxonomy[4] (of 182 concepts nodes).

Figure 4 depicts the visual interface of the CoSeNa system after the user provides the concept "*Iraq*" to start exploration. Coherently to the KbC model, CoSeNa first identifies related content keywords (including "*Saddam Hussein*", "*missile*", "*weapon*", "*Kuwait*", and "*persian gulf* ") and presents these to the user in the form of a keyword cloud.

In addition, using the concept-to-concept similarities (described in Section 4.4), CoSeNa also creates and presents a related concept cloud consisting of geographical concepts "Iran", "United States", "North Korea", and "Russia". These geographical concepts in the concept cloud are also shown on a world map, with markers representing visual links. Note that the CoSeNa interface also shows related videos and images (searched on Youtube and Flickr by using the concept and term clouds) as well as documents that are associated to the concept "Iraq" as described in Section 4.3. When the user clicks on the term, "*weapon*", in the keyword cloud, CoSeNa updates the tag clouds as well as media (text, images, and video) presented to the user accordingly. The result is shown in Figure 5. In this case, the concept cloud (*"Russia", "Iraq", "North*

[3]*http://archive.ics.uci.edu/ml/datasets/Bag+of+Words. This data set has no class labels, and for copyright reasons no filenames or other document-level metadata.*

[4]*The concept taxonomy defines the context that drives the user in searching and navigating the documents. In this case we highlight geographical relationships. The use of a historical taxonomy would instead make evident historical relationship among documents*

Korea", and *"United States"*) represents geographical concepts neighboring the keyword "weapon" in the KbC graph (coherently with the previous case, geographical concepts are shown on the world map). The keyword cloud (*"missile", "security", "arsenal", "warhead"*, etc.) is created using the keyword-to-keyword similarities, as described in Section 4.4.

When the user clicks on a document, as also shown in Figure 5, CoSeNa displays the corresponding article and highlights relevant content and keyword cloud elements in the document.

Note that CoSeNa can use the KbC graph for not only locating the most relevant documents within a given context, but also for (a) explaining why two concepts in the given domain are related and (b) for identifying dominant concepts and keywords in the given context.

5.2 Explaining the Relationships between Concepts in a given Context

Given a document corpus, CoSeNa can *explain* the relationships among the concepts in the associated metadata in terms of keywords extracted from the corpus or explain the relationships among the keywords in terms of the concepts. In order to analyze these semantic relationships between a pair of concepts, we study the combined-vectors, $\vec{clv}$, associated to the nodes corresponding to these concepts in the KbC graph and search for those keyword dimensions (representing the keywords in the KbC graph) which enabled the relationship between these concepts.

More specifically given two taxonomy concepts c_i and c_j , we rank those key-words that occur in both combined-vectors, $\vec{clv_{c_i}}$ and $\vec{clv_{c_j}}$, in terms of their contribution to the relationship between the corresponding concepts. Since the similarity between two combined-vectors are computed based on cosine similarity, we measure the contribution of the keyword *k* to the concepts c_i and c_j as follows:

$$contribution\ (k) = \vec{clv_{c_i}}[k] \,.\, \vec{clv_{c_i}}[k] \qquad (15)$$

After we order the keywords based on their contribution to the relationship of the considered concepts, we select those keywords with the highest contribution.

Note that the process of selecting the concepts that explain the relationship between two keywords is similar: in this case, we analyze the combined-vectors of the keywords nodes in the KbC graph and compute and rank the contributions of the concepts in the combined-vectors.

Note that since the combined-vectors reflect both taxonomical as well as corpus contexts selected by the user, this approach permits understanding the context-specific relationships between two taxonomical concepts or keywords.

5.3 Identifying Dominant Concepts and Keywords in a Given Context

As we mentioned earlier, CoSeNa can leverage the KbC graph for identifying dominant concepts and keywords in a given context. For this, CoSeNa relies on a random-walk based technique that mimics the behavior of a sentient being that navigates over the KbC graph in a way that reflects the strengths of the links. The key observation, also used in web link analysis (Page et al., 1998) and social network analysis (Pujol et al., 2002), is that if this navigation process continues indefinitely, the sentient being will spend most of its time on (concept and keyword) nodes that are strongly linked to the rest of the graph. Therefore, if one can measure the portion of the time the sentient being spends during its infinite random walk on the KbC on a given node, then this can be used to measure the dominance score of the corresponding concept or the keyword.

In CoSeNa, we are relying on a PageRank (Page et al., 1998) like algorithm to compute the dominance scores. More specifically, the authority of a term (a keyword or concept in our KbC graph) depends on the number and the authority of its connected nodes. Hence, given a term $t_i \in (V_C \cup V_K)$, its dominance is defined based on the authorities of its neighbors as follows:

$$dom\ (t_i) = d \times \sum_{t_j \in in(t_i)} \frac{dom\ (t_i)}{dom\ (t_j)} + (1\text{-}d) \qquad (16)$$

where $d \in ((0, 1)$ is a dumping factor which represents the probability with which the sentient being will simply navigate from one node in the KbC graph to another and (*1-d*) represents the probability with which it will jump on an arbitrary node in the graph (*d* is often set to 0.85 (Page et al., [16]), out(*t*) is a function that returns the set of terms that have an incoming edge from

t and in(t) returns the set of terms in the KbC graph that have an outoing edge pointing to t. Since, the definition is recursive, in practice, the dominance values are calculated using an iterative algorithm, where, at the initial instant, each dominance value is initialized to:

$$dom^{o}(t_i) = \frac{1}{Vc \cup Vk} \tag{17}$$

Then, at each step, s, the algorithm recomputes the dominance values based on the dominance scores of the previous step:

$$dom^{s}(t_i) = d \times \sum_{t_j \in in(t_i)} \frac{dom^{s-1}(t_i)}{|out(t_j)|} + (1-d) \tag{18}$$

The process ends when a convergence condition is satisfied. In practice, the convergence is achieved in 9 steps.

6. Experimental Evaluation

In this section, we evaluate the proposed keyword-by-concept graph-based navigation system by analyzing the impact of the parameters introduced before. As mentioned earlier, we used as document corpus the New York Times articles data set (Frank and Asuncion, [9]) (300,000 text entries originally written from 1987 to 2008) and as hierarchy a geographical concept taxonomy with 182 concepts nodes[5]. Thus, we first introduce a quantitative measure in order to evaluate the proposed combined-vector method and we then analyze the impact of the context with real data.

6.1 Measuring the Impact of the Context

As described above, CoSeNa relies on the combined-vectors ($\vec{clv}$) of the concepts and keywords to associate documents to the nodes of the KbC graph. The combined-vectors are also used in determining the strengths of the connections among concepts and among keywords.

As opposed to the concept-vectors (cv), which capture only the taxonomical relationships between concepts, these combined-vectors capture, in addition to the semantic relationships between concepts in the given taxonomy, also the context defined by the documents in the given corpus. In order to observe the impac of this corpus context on the strength of the relationship between a given pair of concepts, c_i and c_j, we define the impact of the corpus context as the ratio:

$$impact(c_i, c_j) = \frac{cos(\vec{clv}_{c_i}, \vec{clv}_{c_i})}{cos(\vec{cv}_{c_i}, \vec{cv}_{c_i})} \tag{19}$$

Figure 5. CoSeNa interface after the selection of keyword "*Weapon*"; in the figure the document visualization interface of CoSeNa which highlights occurrences of the tag cloud terms in the document

[5] *This taxonomy has been extracted from www.dmoz.org/ by using the most cited continent, country and city names in the New York Times corpus.*

Note that if impact(c_i, c_j) ~ 1 then it means that the corpus context has no impact on the strength of the relationship between concepts, c_i and c_j . On the other hand, if impact(c_i, c_j) >>1, then the context defined by the corpus impacts one or both of the concepts in such a way that their relationship strengthens. In contrast, if impact(c_i, c_j) ~ 0, then the impact of the corpus on the concepts, c_i and c_j , is such that their relationship is weakened by the nature of the given set of documents (i.e., the concepts are strongly related to disjoint news events and, thus, the relationship between the concepts is weaker than it is in the given taxonomy).

6.2 Analysis of the Impact of the Context

Table 4(a) shows sample pairs of concepts with most positive, neutral, and most negative impact when using the entire news article corpus. As can be seen here, the content of the news articles significantly strengthens the relationships between concepts, "*Iraq*" and "*United States*", and concepts, "*Europe*" and "*Iran*". In contrast, the relationship between concept pairs, "*Tucson*" and "*London*", has been weakened to almost null. In fact, the keyword clouds corresponding to these two concepts show that, while the former is related to immigration news (with keywords such as "*border patrol*" and "*u.s. border*"), the latter is highly related to sports and arts news (with keywords, such as "*hamilton*" –the name of a British Formula1 driver–, "*spectator*", "*art*", and "theater").

Table 4(b), on the other hand, shows sample pairs of concepts with most positive, neutral, and most negative impact when the set of documents used for evidence-vector computation are limited to those containing the keyword "*economy*". As can be seen here, the content of the economy related news articles significantly strengthens the relationships between geographic concepts pairs, "*United States*"-"*China*", "*United States*"-"*Japan*","*United States*"-"*Taiwan*" and "*Europe*"-"*Russia*". It is

Concept 1	Concept 2	Impact
Cuba	Florida	71.60 (Strengthened)
Europe	Iran	55.61 (Strengthened)
Iraq	United States	48.51 (Strengthened)
Afghanistan	United Stated	29.27 (Strengthened)
...	...	...
North America	United Stated	1.01 (No impact)
Las Vegas	Nevada	0.99 (No impact)
...	...	...
Madrid	Houston	~ 0 (Weakened)
London	Tucson	~ 0 (Weakened)

(a) Using all the available news articles

Concept 1	Concept 2	Impact
United States	China	68.28 (Strengthened)
United States	Japan	43.12 (Strengthened)
United States	Taiwan	41.28 (Strengthened)
Europe	Russia	21.24 (Strengthened)
...	...	...
South America	Brazil	1.01 (No impact)
North America	Canada	0.99 (No impact)
...	...	...
New York	Harare	~ 0 (Weakened)
Paris	Sydney	~ 0 (Weakened)

(b) Using the "economy" articles

Table 4. The impact of the corpus context: example (a) relationships that are strengthened and weakened using the context defined by the entire corpus of news articles; example (b) relationships that are strengthened and weakened using the context defined by the news articles containing the term "economy".

important to note that, as expected, the sets of concept pairs that are most positively and most negatively impacted (i.e., strengthened and weakened) are different when the user focus is different.

6.3 Analysis of the Concept-Keyword Relationships

As explained in Section 5.2, CoSeNa can use the KbC graph for explaining why two concepts or keywords in the given domain are related in a given context. Table 5 reports the most relevant terms that caused the strengthening of the relationships between the geographic concepts pairs reported in Table 4; as explained in Section 5.2, the listed keywords are shared by the combined-vectors of the pair of concepts, and contributed strongly to the creation of the semantic links in the KbC graph (Section 3.5). For example, considering the geographical concepts "*Florida*" and "*Cuba*", the strenght of this semantic relationship is based on terms as "*Elian Gonzalez*" – a young Cuban boy who was at the center of a controversy involving the governments of Cuba and the United States –, "*immigration*", "*Naturalization Service*" and "*Miami*" that help define the nature of the relationship.

On the other hand, when we focus on the KbC graph created based on the subset of documents related to "economy", it is possible to note that the terms that caused the strengthening of the relationships among geographical concepts are strictly

Concept 1	Concept 2	Keywords
Cuba	Florida	Elian Gonzalez, immigration, Naturalization Service, Miami
Europe	Iran	security, Middle East, intelligence, petroleum
Iraq	United States	weapon, defense, gulf war, inspection
Afghanistan	United Stated	taliban, terrorism, Bin Laden, Pakistan

(a) Using all the available news articles

Concept 1	Concept 2	Keywords
United States	China	World Trade Organization, Clinton, globalization, cooperation
United States	Japan	technology, consumer, investor, sony
United States	Taiwan	independence, threat, negotiation, tension
Europe	Russia	agreement, energy, Ukraine, crisis

(b) Using the "*economy*" articles

Table 5. The most relevant terms, in the combined-vectors, that guide the strength-ening of (a) the relationships between concepts using the context defined by the entire corpus of news articles and (b) the relationships between concepts using the context defined by the news articles containing the term "*economy*".

related to the economic domain. For example, the relationships between "United States" and "China" has been strenghthened based on terms as "World Trade Organization" or "globalization" that represent highly focussed keywords in the considered domain. Therefore, the KbC graph reflects the context in which it is created and can be used for explaining the relationships between the concepts of interest within the given context.

6.4 Analysis of the Keyword and Concept Dominance Scores

As described in Section 5.3, CoSeNa can also use the KbC graph for identifying dominant concepts and keywords in a given context.

In Tables 6(a) and (b) the most dominant keywords and concepts based on the full document set are reported. As can be seen in 6(a), when the full data set is considered most of dominant terms coming from the considered corpus (Table 6(a)) are political (names of american politicians or general terms semantically related to politics) and technology related. Similarly, the most dominant geographical concepts (Table 6(b)) represent the countries (such as "*Israel*", "Lebanon", "*China*", "*Russia*", "Cuba", "*Colombia*", and "*Iraq*") and cities (such as "*Miami*" –which is related to "*Cuba*"– and "*New York*") related to foreign political and economic relationships of the United States in the period covered by the selected corpus.

In contrast, in Tables 7(a) and (b) the most dominant keywords and concepts based on the subset of corpus containing the

term “economy” are reported. As can be seen in these tables, when only the economy related subset is considered, the most dominant keywords and the most dominant concepts reflect the considered corpus context and highlight the economic focus of this subset of news articles.

7. Conclusions

In this paper, we proposed a novel keywords-by-concepts (KbC) graph, which is a weighted graph constructed relying on a spreading activation like technique by a tight integration of the available domain taxonomies (considered as the semantic

Ranking	Keyword Node	Dominance value
1st	Al Gore	0.0032
2nd	George Bush	0.0028
3th	John McCain	0.0015
4th	government	0.0013
5th	computer	0.0012
6th	internet	0.0011
7th	voter	0.0010
8th	democrat	9.7403E-4
9th	woman	9.1351E-4
10th	technology	7.6794E-4

(a) Node of KbC graph from the corpus
(using all the available articles)

Ranking	Concept Node	Dominance value
1st	Israel	5.6029E-4
2nd	Miami	4.4354E-4
3th	Lebanon	3.6508E-4
4th	United States	3.4079E-4
5th	China	3.1240E-4
6th	Russia	2.5890E-4
7th	Cuba	2.2968E-4
8th	New York	1.4158E-4
9th	Colombia	9.7460E-5
10th	Iraq	9.2171E-5

(b) Node of KbC graph from the *taxonomy*
(using all the available articles)

Table 6. (a) The most dominant keyword nodes on the KbC graph generated using the entire New York Times corpus and (b) the most dominant concept nodes on the KbC graph generated using the same document corpus.

context) with the keywords extracted from the available corpus of documents. KbC graph is then leveraged for developing a novel Context-based Search and Navigation (CoSeNa) system for context-aware navigation and document retrieval. The unique aspect of our approach is that it mines emerging topic correlations within the data, exploiting both statistical information coming from the document corpus and the structured knowledge represented by the input taxonomy. The case study,

Ranking	Keyword Node	Dominance value
1st	George Bush	0.0027
2nd	Al Gore	0.0026
3th	percent	0.0020
4th	government	0.0016
5th	White House	0.0013
6th	statesman	0.0012
7th	United States	0.0012
8th	business	9.3666E-4
9th	.com	7.7473E-4
10th	economy	7.7328E-4

(a) Node of KbC graph from the corpus (using all the available articles)

Ranking	Concept Node	Dominance value
1st	United States	0.0012
2nd	China	6.6503E-4
3th	Israel	4.9638E-4
4th	Russia	3.4715E-4
5th	Mexico	2.6899E-4
6th	Japan	2.6759E-4
7th	New York	2.6169E-4
8th	Argentina	2.1334E-4
9th	Zimbabwe	1.6380E-4
10th	Taiwan	1.5477E-5

(b) Node of KbC graph from the taxonomy (using all the available articles)

Table 6. (a) The most dominant keyword nodes on the KbC graph generated using the entire New York Times corpus and (b) the most dominant concept nodes on the KbC graph generated using the same document corpus

presented in the paper, shows how this approach enables contextually-informed strengthening and weakening of semantic links between different concepts.

References

[1] Aswath, Dipti., Toufeeq, Syed Ahmed, cunha, James D'., Davulcu, Hasan (2005). Boosting item keyword search with spreading activation. *In:* WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, p. 704–707, Washington, DC, USA, 2005. IEEE Computer Society.

[2] Bates, Marcia J. (1989). The design of browsing and berrypicking techniques for the online search interface, *Online Review*, 13 (5) 407–424.

[3] Crestani, F. (1997). Application of spreading activation techniques in information retrieval, *Artificial Intelligence Review*, 11. 453–482.

[4] Cui, Hang., Wen,Ji-Rong., Nie, Jian-Yunand., Ma,Wei-Ying. (2002). Probabilistic query expansion using query logs, *In:* WWW '02. Proceedings of the 11th international conference on World Wide Web, p. 325–332, New York, NY, USA. ACM.

[5] Deerwester, Scott., Dumais, Susan T., Furnas, George W., Landauer, Thomas K., Harshman, Richard (1990). Indexing by latent semantic analysis, J*ournal of the American Society for Information Science*, 41. 391–407.

[6] Eckart, Carl., Young, Gale. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1 (3) 211–218.

[7] Ellis, D. (1998) . A behavioral approach to information retrieval system design. *J. Doc.*, 45 (3) 171–212.

[8] Fellbaum.(1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

[9] Frank,A., Asuncion, A. (2010). UCI machine learning repository. URL http://archive.ics.uci.edu/ml.

[10] Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A.,Streeter, L. A., Lochbaum,K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. *In SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–480, New York, NY, USA, 1988. ACM.

[11] Gauch, Susan., Chaffee, Jason., Pretschner, Alaxander (2003). Ontology-based personalized search and browsing. *Web Intelli. and Agent Sys.*, 1(3-4) 219–234, 2003.

[12] Grootjen,F. A., van der Weide,Th. P. (2006). Conceptual query expansion. *Data Knowl. Eng.*, 56 (2) 174–193.

[13] Kim, Jong Wook., Selcuk Candan, k (2006). Cp/cv: concept similarity mining without frequency information from domain describing taxonomies. *In:* CIKM '06, p. 483–492.

[14] Li,Wen-Syan., Selcuk Candan, K. (1999). Semcog: A hybrid object-based image and video database system and its modeling, language, and query processing. *TAPOS*, 5 (3) 163–180.

[15] Mandala, Rila., Tokunaga, Takenobuand., Tanaka, Hozumi.(1999). Combining multiple evidence from different types of thesaurus for query expansion. *In:* Proc of ACM SIGIR'99, p. 191–197.

[16] Page, L.S., Brin, R., Motwani, Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. *In:* WWW'98, p.161–172.

[17] Preece, Scott Everett (1981). A spreading activation network model for information retrieval. PhD thesis, Champaign, IL, USA.

[18] Pujol, Josep M., Sang¨uesa, Ramonand Delgado,Jordi (2002). Extracting reputation in multiagent systems by means of social network topology. *In*: AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems, p. 467–474, New York, NY, USA. ACM.

[19] Qiu,Yonggang., Frei,Hans-Peter. (1993). Concept based query expansion. *In:* SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, p. 160–169, New York, NY, USA, 1993. ACM.

[20] Rieh, Soo Young., Xie,Hong (2006). Analysis of multiple query reformulations on the web: the interactive information retrieval context. *Inf. Process. Manage.*, 42 (3) 751–768.

[21] Ruthven, Ian., Lalmas, Mounia (2003). A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18 (2) 95–145.

[22] Maria Sacco, Giovanni.(2000). Dynamic taxonomies: A model for large information bases, *IEEE TKDE*, 12 (3) 468–479 .

[23] Salton,Gerard., Buckley, Christopher (1998). Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, 513–523.

[24] Pandit, Shashank (1998). Navigation-aided retrieval. *In:* Proc of WWW'07, p. 391–400.

[25] Jones, Karen Sparck (1991). Notes and references on early automatic classification work. SIGIR Forum, 25 (1) 10–17.

[26] Spink, Amanda., Building, Rider I., Wolfram, Dietmar., Saracevic,Tefko (2001). Searching the web: the public and their queries. *J. of the American Society for Information Science and Technology*, 52. 226–234.

[27] Teevan, Jaime., Alvarado, Christine., Ackerman, Mark S., Karger,David R.(2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. *In:* SIGCHI'04, p. 415–422. ACM.

[28] Xu, Jinxi., Croft,W. Bruce (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18 (1) 79–112.

Author biographies

Claudio Schifanella received his Ph.D. degree in Computer Science in 2008 at the University of Torino, where he is currently a short term researcher. His research topics involve data mining on digital documents and unsupervised learning.

Mario Cataldi got his M.S. and Ph.D. degrees in Computer Science at the University of Torino, where he is currently Post Doctoral researcher. His current research interests include text mining, data summarization and knowledge management. His active collaborations include RAI CRIT, Telecom Italia lab. in Torino and the Arizona State University.

K. Selcuk Candan is a Professor of Computer Science and Engineering at the Arizona State University. He joined ASU after receiving his Ph.D. from the University of Maryland at College Park. Prof. Candan's primary research interest is in the area of management of non-traditional, heterogeneous, and imprecise (such as multimedia, web, scientific) data. Prof. Candan is an editorial board member of the Very Large Databases (VLDB) journal. He is also in the editorial board of the Journal of Multimedia. He has served in the organization and program committees of various conferences (SIGMOD06, ACM MM08, SIGMOD10, ACM MM11,SIGMOD12)

Maria Luisa Sapino got her PhD degree in Computer Science at the University of Torino, where she is currently a Professor. Her initial scientific contributions were in the area of logic programming and artificial intelligence, specifically in the semantics of negation in logic programming, and in the abductive extensions of logic programs. Since mid-90s she has been applying these techniques to the challenges associated with database access control, and with heterogeneous and multimedia data management. In particular, she developed novel techniques and algorithms for similarity based information retrieval, content based image retrieval, web accessibility for users who are visually impaired.

Luigi Di Caro got his PhD degree in Computer Science at the University of Torino. His contribution mostly involved Text Mining and Data Visualization techniques.In particular, he developed different algorithms and systems to explore text corpora as well as methods to track the evolution of topics over time. His current research interests are mostly addressed towards the integration of statistical analysis and Natural Language Processing techniques.