CoSeNa: a Context-based Search and Navigation System

Mario Cataldi
Università di Torino
Torino, Italy
cataldi@di.unito.it

Claudio Schifanella
Università di Torino
Torino, Italy
schi@di.unito.it

K. Selçuk Candan
Arizona State University
Tempe, AZ 85283, USA
candan@asu.edu

Maria Luisa Sapino
Università di Torino
Torino, Italy
mlsapino@di.unito.it

Luigi Di Caro
Università di Torino
Torino, Italy
dicaro@di.unito.it

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. In this paper, we also present CoSeNa (*Context-based Search and Navigation*) system that leverages the KbC model as the basis for document exploration and retrieval as well as contextually-informed media integration.

Categories and Subject Descriptors

H.5.4 [**Hypertext/Hypermedia**]: Navigation; I.7.3 [**Document and Text Processing**]: Index Generation; H.3.3 [**Information Search and Retrieval**]: Information Filtering, Selection Process; H.3.7 [**Digital Libraries**]: System Issue

General Terms

Algorithms, experimentation

Keywords

Web based Digital Ecosystems; Data Knowledge Management; HCI

MEDES 2009 October 27-30, 2009, Lyon, France

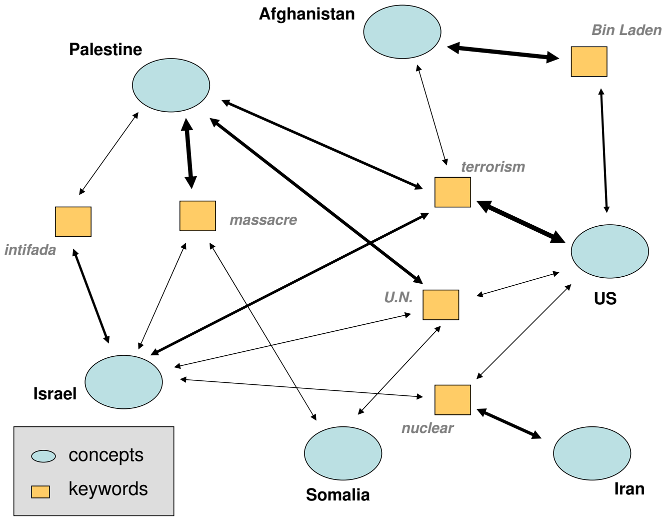


Figure 1: An example KbC graph constructed using concepts from a geographical domain taxonomy and keywords extracted from a corpus of news documents

1. INTRODUCTION

Popular approaches to text retrieval are mostly based on available feature statistics [13]. Some recent systems also leverage available semantics to guide the retrieval process towards an equilibrium between relatedness and wisdom [4]. 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

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.

1.1 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.

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 [15]. 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. [8] presents 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. In [9] the authors present a method for expanding target concepts of the whole query instead of a term-by-term change. A well known query reformulation method is *user relevance feedback* [11]. The idea is to ask the users to mark relevant documents in a 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 [5], 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 [3]. However, these schemes still assume that user's initial query is highly precise and its expansion is sufficient to identify the relevant documents.

An alternative approach to retrieval is to rely on an exploratory process instead of document indexing and query matching [7]. As stated in [14] 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 [10] 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 [16].

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 [2, 14] help address the "don't-know-what-I-want" behavior [1] more effectively than relevance feedback schemes that assume that the user knows what she wants.

1.2 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 [12]. 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

the keywords appearing in the news articles.

[2] We will discuss in Section 2.5 how these content keywords are extracted and how the links are defined

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) (Figure 1);

- 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;

- 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.

The rest of the paper is organized as follows: Section 2 reports the algorithm used to define the semantic correlations among concepts and keywords and explains how to build a KbC navigational graph. Section 3 defines how to bind the most relevant documents to each graph node, using the semantic information inferred by the graph. Section 4 shows the features of the implemented system.

2. KEYWORDS-BY-CONCEPTS GRAPH CONSTRUCTION

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 4, 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-step 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. Step 1 preprocesses the corpus of documents, to extract keyword frequencies (Section 2.1).

2. The second 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 2.2).

3. For each document in the database, 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 2.3).

4. For each concept, step 4 extracts the most relevant keywords contained in the documents described under it (Section 2.4). This helps identify highly correlated concepts and keywords, providing the basis for the *keywords-by-concepts* (KbC) navigational graph construction.

Next, we discuss these steps in detail.

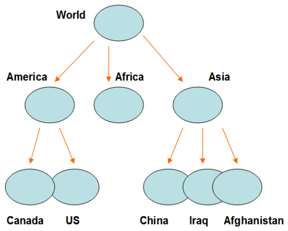
2.1 Text Analysis: Keyword Vector Extraction

As in most IR systems, the analysis process starts with extraction of the keywords from the given corpus. The corpus, D, of the content documents is analyzed and a representative keyword vector is generated for each document. The $m = |D|$ documents (i.e., articles) are represented with vectors in which each component represents a keyword. As usual, the keyword extraction includes a preliminary phase of stop word elimination and stemming [3]. The weight associated to each stemmed term is computed in the augmented normalized term frequency form [13]. For the i^{th} corpus document, a keyword vector $\vec{d_i} = \{w_{i,1}, w_{i,2}, ..., w_{i,v}\}$ is defined, where v is the size of the considered vocabulary, $1 \leq i \leq m$, and $w_{i,j}$ is the normalized term frequency of the $j-th$ vocabulary term in the $i-th$ document.

2.2 Taxonomy Analysis: 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, $\mathcal{H}(C, E)$, onto a concept-vector space. More specifically, given a taxonomy, $\mathcal{H}(C, E)$, with $n = |C|$ concepts, we represent each concept node as a vector $\vec{cv}$ with n dimensions

[3] For stemming, we rely on *Wordnet* [3]. Stemmed entries are associated to keyword vectors.



	world	Asia	Africa	America	Afghanistan	Iraq	China	Canada	US
$\vec{cv}_{world}$	0.450	0.169	0.141	0.158	0.018	0.018	0.018	0.021	0.021
$\vec{cv}_{Asia}$	0.052	0.469	0.006	0.006	0.156	0.156	0.156	0.0003	0.0003
$\vec{cv}_{Africa}$	0.100	0.012	0.873	0.012	0.0006	0.0006	0.0006	0.0007	0.0007
$\vec{cv}_{America}$	0.057	0.007	0.007	0.520	0.0003	0.0003	0.0003	0.204	0.204
$\vec{cv}_{Afghanistan}$	0.004	0.100	0.0002	0.0002	0.872	0.012	0.012	0	0
$\vec{cv}_{Iraq}$	0.004	0.100	0.0002	0.0002	0.012	0.872	0.012	0	0
$\vec{cv}_{China}$	0.004	0.100	0.0002	0.0002	0.012	0.012	0.872	0	0
$\vec{cv}_{Canada}$	0.006	0.0003	0.0003	0.165	0	0	0	0.806	0.023
$\vec{cv}_{US}$	0.006	0.0003	0.0003	0.165	0	0	0	0.023	0.806

Table 1: A geographical taxomomy and the related concept vectors

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 in [6]. 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 concept vectors are obtained by propagating concepts on the taxonomy graph according to their *semantic* contributions (dictated by the structure of the taxonomy).

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

$$\langle 0.450, 0.169, 0.141, 0.158, 0.018, 0.018, 0.018, 0.021, 0.021\rangle,$$

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".

2.3 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 is *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 dot product 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 $d_j \in D$:

1. consider the document vector $\vec{dv}_j$

2. compute its similarity wrt. all the concept vectors describing the given taxonomy.

$$sim(\vec{cv}_i, \vec{dv}_j) = \Sigma_{k=1}^{u} \vec{cv_i}[k] \times \vec{dv_j}[k]$$

3. sort the concepts 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 $\vec{dv}_j$, 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 nonempty set of concepts associated to it.

2.4 Discovery of Concept-Keyword Mappings

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 $D_{desc}(c\vec{v}_{c_i})$. Notice that, in general, the sets of associated document 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 not be associated as a descriptive concept to any of the documents in the database. For such a concept, the corresponding set, D_{desc}, of associated documents is empty.

At this step, given a concept c_i and the set, $D_{desc}(c\vec{v}_{c_i})$, 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 $D_{desc}(c\vec{v}_{c_i})$ 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 to the 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 relevance feedback on the results of such query. Recognizing this, given a concept c_i and a corresponding set of associated documents, $D_{desc}(c\vec{v}_{c_i})$, we identify the weight, $u_{i,j}$, of the keyword k_j relying on a probabilistic feedback mechanism [11]:

$$u_{i,j} = log\frac{r_{i,j}/(R_i - r_{i,j})}{(n_j - r_{i,j})/(N - n_j - R_i + r_{i,j})} \times \left| \frac{r_{i,j}}{R_i} - \frac{n_j - r_{i,j}}{N - R_i} \right|,$$

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 $D_{desc}(c\vec{v}_{c_i})$; 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*, $\vec{lv}_{c_i}$.

2.5 Constructing the KbC Graph using the Concept-Keyword Mappings

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

$$\vec{lv}_{c_i} = \langle u_{i,1}, u_{i,2}, \ldots, \rangle,$$

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 it at least one evidence vector. We construct KbC as follows 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, $V_C = \{v_{c_1}, \ldots, v_{c_n}\}$, where vertex $v_{c_i} \in V_C$ is labelled as "c_i"; i.e., $l(v_{c_i})$ ="c_i";
- Let V_K be a set of vertices, $V_K = \{v_{k_1}, \ldots, v_{k_m}\}$, where vertex $v_{k_j} \in V_K$ is 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 $\vec{lv}_i[j] \neq 0$, there exists an edge $\langle v_{c_i}, v_{k_j} \rangle \in E$ such that

$$\rho(\langle v_{c_i}, v_{k_j} \rangle) = \rho_{i,j} = \frac{\vec{lv}_{c_i}[j]}{\left\| \vec{lv}_{c_i} \right\|}$$

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

3. UNIFYING CONCEPT AND KEYWORD VECTOR SPACES TO SUPPORT DOCUMENT RETRIEVAL

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 2.3, 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 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.

3.1 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,

$$\vec{clv}_{c_i} = \alpha_{c_i} \cdot \vec{cv}_{c_i} + \beta_{c_i} \cdot \vec{lv}_{c_i},$$

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 2.3, 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, α_{c_i} and β_{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

$$\frac{\|\alpha_{c_i} \cdot \vec{cv}_{c_i}\|}{\left\|\beta_{c_i} \cdot \vec{lv}_{c_i}\right\|} = \frac{|C_{c_i}|}{|L_{c_i}|}.$$

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

$$\frac{\alpha_{c_i}}{\beta_{c_i}} = \frac{|C_{c_i}|}{|L_{c_i}|}.$$

Also, if we further constrain that the combined vector $\vec{clv}_{c_i}$ is also normalized to 1,

$$\left\|\alpha_{c_i} \cdot \vec{cv}_{c_i} + \beta_{c_i} \cdot \vec{lv}_{c_i}\right\| = 1,$$

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

$$\alpha_{c_i} = \frac{|C_{c_i}|}{|C_{c_i}| + |L_{c_i}|} \quad \text{and} \quad \beta_{c_i} = \frac{|L_{c_i}|}{|C_{c_i}| + |L_{c_i}|}.$$

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

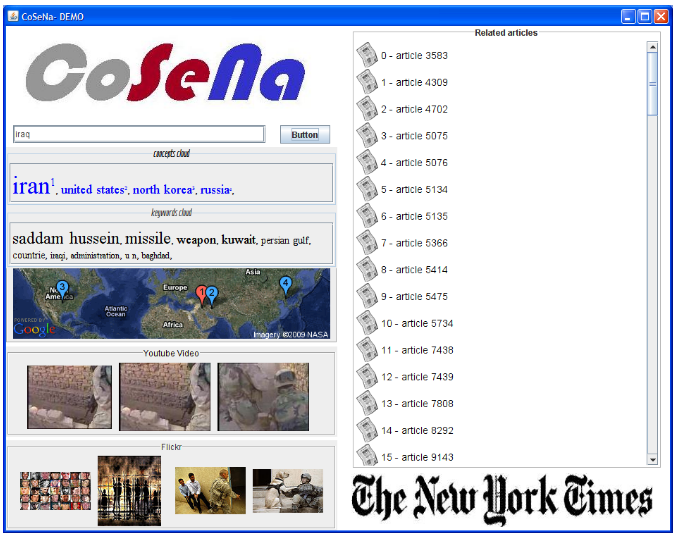
$$\vec{clv}_{c_i} = \frac{|C_{c_i}|}{|C_{c_i}| + |L_{c_i}|} \cdot \vec{cv}_{c_i} + \frac{|L_{c_i}|}{|C_{c_i}| + |L_{c_i}|} \cdot \vec{lv}_{c_i}.$$

3.2 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 $\vec{clv}_{k_j}$ is computed as

$$\vec{clv}_{k_j} = \sum_{c_i \in neighbor(v_{k_j})} \left(\frac{\rho_{i,j}}{\left\|\vec{lv}_{c_i}\right\|} \cdot \vec{lv}_{c_i} \right),$$

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 2.5. As it is the case for the $\vec{clv}_{c_i}$ vectors, $\vec{clv}_{k_j}$ are also normalized to 1.



iran[1], united states[2], north korea[3], russia[4],

saddam hussein, missile, weapon, kuwait, persian gulf, countrie, iraqi, administration, u n, baghdad,

0 - article 3583
1 - article 4309
2 - article 4702
3 - article 5075
4 - article 5076
5 - article 5134
6 - article 5135
7 - article 5366
8 - article 5414
9 - article 5475
10 - article 5734
11 - article 7438
12 - article 7439
13 - article 7808
14 - article 8292
15 - article 9143

Figure 2: CoSeNa search with geographical concept "Iraq"

3.3 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 $\vec{clv}$, the documents in the given corpus can be associated under these nodes as in Section 2.3, but using $\vec{clv}$ vectors instead of $\vec{cv}$ 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.

3.4 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 $\vec{clv}$, capturing both the taxonomical relationships between concepts and the context defined by the documents in the given corpus. Therefore, in addition 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 the next section, we will describe use of this in CoSeNa to support document exploration.

4. COSENA SYSTEM AND USE CASE

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 preprocessing 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).

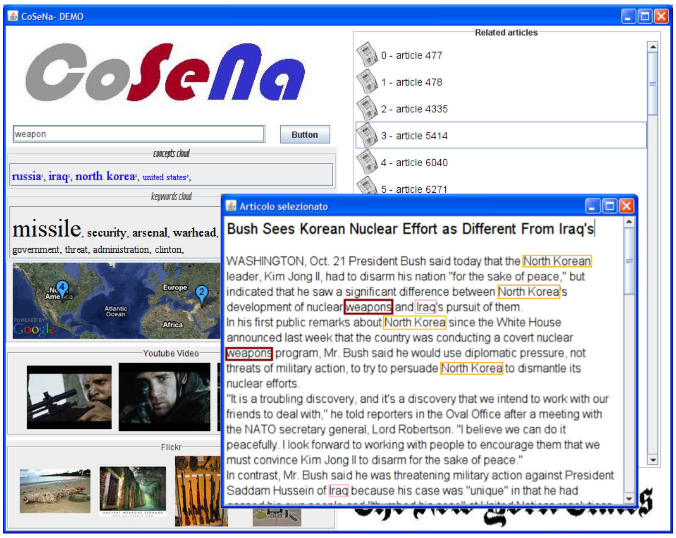
Concept and keyword font sizes express the strength of the relationships among concepts and keywords. Documents associated to the concepts and keywords are enumerated in a list ordered with respect to the weights calculated in Section 3.3. When the user clicks on a document, the system shows the corresponding document and highlights the contextually important concepts and keywords in the document. The user can navigate into the KbC space by clicking on the concepts and keywords highlighted in the tag clouds as well as in the documents.

CoSeNa also provides media integration with three online sources: Google Maps, Flickr, and YouTube. To achieve context-based integration, CoSeNa queries the content sources using the concepts and keywords in the clouds and presents the results to the user in a unified interface.

4.1 Navigational Interface

Figures 2 and 3 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[4]) is explored with the help of a geographical concept taxonomy[5] (of 182 concepts nodes).

Figure 2 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 3.4), 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 3.3. 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 3. In this case, the concept cloud ("russia", "iraq", "north 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 3.4. When the user clicks on a document, as also shown in Figure 3, CoSeNa displays the corresponding article and highlights relevant content and keyword cloud elements in the document.



Bush Sees Korean Nuclear Effort as Different From Iraq's

WASHINGTON, Oct. 21 President Bush said today that the North Korean leader, Kim Jong Il, had to disarm his nation "for the sake of peace," but indicated that he saw a significant difference between North Korea's development of nuclear weapons and Iraq's pursuit of them.
In his first public remarks about North Korea since the White House announced last week that the country was conducting a covert nuclear weapons program, Mr. Bush said he would use diplomatic pressure, not threats of military action, to try to persuade North Korea to dismantle its nuclear efforts.
"It is a troubling discovery, and it's a discovery that we intend to work with our friends to deal with," he told reporters in the Oval Office after a meeting with the NATO secretary general, Lord Robertson. "I believe we can do it peacefully. I look forward to working with people to encourage them that we must convince Kim Jong Il to disarm for the sake of peace."
In contrast, Mr. Bush said he was threatening military action against President Saddam Hussein of Iraq because his case was "unique" in that he had

Figure 3: 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

4.2 Contextual Impact

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 ($\vec{cv}$), 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 impact 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

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

Note that if $impact(c_i, c_j) \sim 1$, 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 $impact(c_i, c_j) \gg 1$, 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 $impact(c_i, c_j) \sim 0$, 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 document (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).

Table 2(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 strengthen 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

[4] 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.

[5] 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

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 2: 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".

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 2(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 strengthen the relationships between geographic concepts pairs, "United States"-"China", "United States"-"Japan","United States"-"Taiwan" and "Europa"-"Russia". It is 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.

5. CONCLUSIONS

In this paper, we propose a novel *keywords-by-concepts* (KbC) graph, which is a weighted graph constructed by a tight integration of the available domain taxonomies (considered as the semantic context) with the keywords extracted from the available corpus of documents. KbC graph is then leveraged for developing a novel *a 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, presented in the paper, shows how this approach enables contextually-informed strengthening and weakening of semantic links between different concepts.

6. REFERENCES

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

[2] D. Ellis. A behavioral approach to information retrieval system design. *J. Doc.*, 45(3):171–212, 1989.

[3] Fellbaum. *WordNet: An Electronic Lexical Database.* The MIT Press, May 1998.

[4] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. *Web Intelli. and Agent Sys.*, 1(3-4):219–234, 2003.

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

[6] J. W. Kim and K. S. Candan. Cp/cv: concept similarity mining without frequency information from domain describing taxonomies. In *CIKM '06*, pages 483–492, 2006.

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

[8] R. Mandala, T. Tokunaga, and H. Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *Proc of ACM SIGIR'99*, pages 191–197, 1999.

[9] Y. Qiu and H.-P. Frei. Concept based query expansion. In *SIGIR '93*, pages 160–169. ACM.

[10] S. Y. Rieh and H. Xie. Analysis of multiple query reformulations on the web: the interactive information retrieval context. *Inf. Process. Manage.*, 42(3):751–768, 2006.

[11] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, June 2003.

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

[13] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.

[14] S. P. Shashank. Navigation-aided retrieval. In *Proc of WWW'07*, pages 391–400.

[15] A. Spink, R. I. Building, D. Wolfram, and T. Saracevic. Searching the web: the public and their queries. *J. of the American Society for Information Science and Technology*, 52:226–234, 2001.

[16] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *SIGCHI'04*, pages 415–422. ACM, 2004.