Knowl Inf Syst (2014) 38:109–140
DOI 10.1007/s10115-012-0493-x

REGULAR PAPER

Context-based information analysis for the Web environment

Vesile Evrim · Dennis McLeod

Received: 18 November 2010 / Revised: 25 December 2011 / Accepted: 18 March 2012 /
Published online: 30 March 2013

Abstract Finding the relevant set of information that satisfies an information request of a Web user in the availability of today's vast amount of digital data is becoming a challenging problem. Currently, available Information Retrieval (IR) Systems are designed to return long lists of results, only a few of which are relevant for a specific user. In this paper, an IR method called Context-Based Information Analysis (CONIA) that investigates the context information of the user and user's information request to provide relevant results for the given domain users is introduced. In this paper, relevance is measured by the semantics of the information provided in the documents. The information extracted from lexical and domain ontologies is integrated by the user's interest information to expand the terms entered in the request. The obtained set of terms is categorized by a novel approach, and the relations between the categories are obtained from the ontologies. This categorization is used to improve the quality of the document selection by going beyond checking the availability of the words in the document by analyzing the semantic composition of the mapped terms.

Keywords Information retrieval · Ontology · Context-based search · Relevance · Query

V. Evrim (✉)
Information System Engineering,
Cyprus International University,
via Mersin 10, Haspolat-Lefkoşa, North-Cyprus, Turkey
e-mail: vesile@gmail.com

D. McLeod
Semantic Information Research Laboratory,
Department of Computer Science,
University of Southern California,
Los Angeles, CA, USA
e-mail: mcleod@usc.edu

1 Introduction

As the amount of available digital information grows, the ability to search for information becomes increasingly critical. Document retrieval on the World Wide Web (WWW), with over 29 billon pages [39], is a challenging task for more than 1.5 billion users [34]. The information providers in the Web follow few formal protocols, often remain anonymous and publish in a wide variety of formats. There is no central registry to access the content of WWW, and the documents are often misrepresent their content while some authors try to change the ranking of the documents for their personal favor [21].

In order to retrieve the information, WWW users use search engines. Currently, the most popular search engines used by the Web users to retrieve the information are Google and Yahoo [34]. A widely accepted de-facto standard for Web search is a simple query form that accepts keywords and returns as a result document locators (URLs). In general, keyword-based query articulation is difficult. Therefore, typical queries are short, comprising an average of two to three terms per query [17,48].

Given the user query, the key goal of an Information Retrieval (IR) System is to retrieve information that might be useful or relevant to the user. However, inferring user preferences from a few keywords is a difficult task. In fact, most state-of-the-art retrieval models ignore this problem altogether and simply treats queries and documents as a bag of words. Therefore, the need of context investigation becomes necessary [53]. Baralis et al. introduced a CAS-MINE framework to efficiently discover relevant relationships between user context data and currently asked services for both user and service profiling.

The user with an information request can be anybody and can search about anything. Broader [6] identified the needs of the users search on the Web as Navigational, Informational and Transactional. In general, it is really hard to determine the intent of a user since one user can have different goals in different searches or can combine two or more search needs to satisfy a single goal.

Rose and Levinson [43] divided informational search to five subcomponents: "directed" (open, closed), "undirected," "find," "list" and "advice." While "directed" search is defined for searching a specific answer, "undirected" search is defined for searching all the information about the topic.

The quality of the returned documents becomes even more important when the user is a domain expert and the information is going to be used for critical decisions such as in Terrorism and Health domains. The need of new Information Retrieval methodology for the users of these domains is beyond Navigational and Transactional uses. The dynamic structure of Web makes necessary for these users to retrieve dynamic information to better understand the up-to-date changes.

Customizing the search results for the critical domain users is the subject of this paper. The previous categorizations made about the search intent do not show the real intent of the users in these domains. In order to better customize the definition of "informational" search within the context of this paper, we further subcategorized "informational" search as "Static" and "Dynamic."

"Static Informational" search is defined as an intent of a user to find the definition or description information about a given topic. On the other hand, "Dynamic Informational" search is defined as an intent of user to find the event-related information about a given topic. The information requested in "Dynamic Information" search category is not a unique answer but a set of answers that will be a response to the user's request.

Given a domain user with a dynamic information need, the objective of this research is to analyze the domain and the information requests of the user, by using semantic

information, to retrieve the most relevant set of documents to satisfy a user's request. In Sect. 2, the architecture of the system, used ontologies and WordNet, is explained. Section 3 describes the extension of the query words and how the document relevance is calculated. Finally, Sect. 4 presents the result of the experiments that is used to test the performance of CONIA.

2 System architecture

This paper presents a novel idea to return the best possible set of documents in response to information need expressed by a user through an inputted query. CONIA is not a crawler; thus, it does not crawl the Web to find the relevant documents for the user's query. Instead, it uses other search engines' returned results for the user's query to build up the corpus.

In order to provide the best possible set of documents in response to a user's request, CONIA combines different components together. The flow of the architecture can be described in Fig. 1. The query entered by the user passes to the search engine such as Google/Yahoo for the document retrieval. The returned documents as response to the query form the corpus that CONIA will use in ranking. Besides, the query also passes to the Ontol-

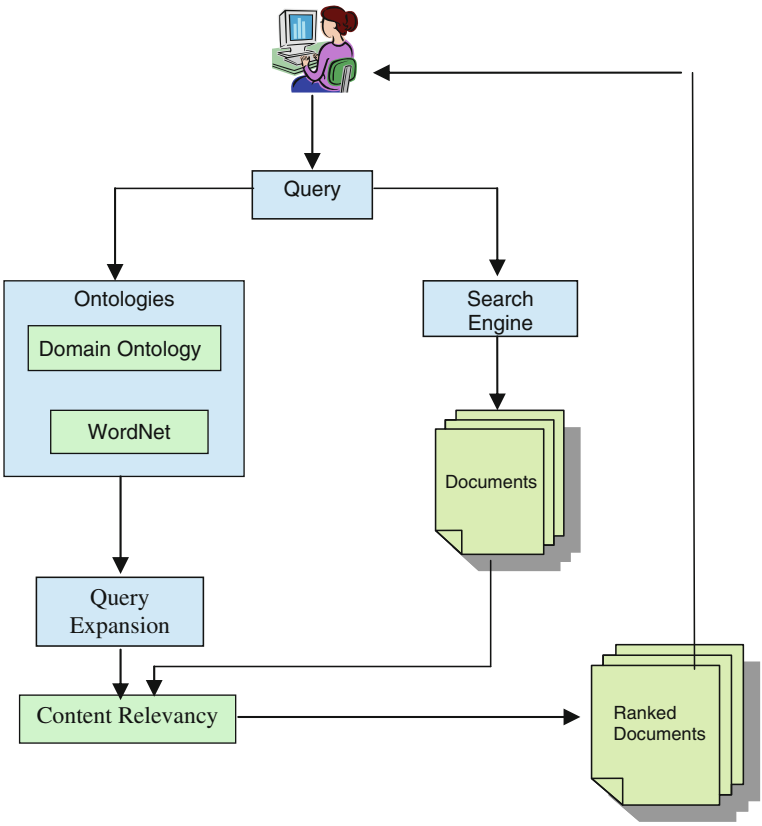

Fig. 1 CONIA framework

ogies for expansion and semantic analysis purposes. Finally, the expanded query terms and the corpus of the documents pass to the "Content Relevancy" module to be used in the computation of relevancy value that will be used in the ranking of the documents as described in details in the following sections.

2.1 Query expansion by using WordNet

The number of words used in a Google query to search for information increased from 3 to 4 words in 2008 [51]. This is a clear indication that users desire to get more relevant result by expressing their need with more words. Different query expansion techniques [27,46] are used to expand the query in mapping the documents.

WordNet, as the biggest lexical handmade ontology, has become one of the most practical taxonomies used in query expansion. Chen et al. applied WordNet to discover fuzzy frequent itemsets as candidate cluster labels for grouping documents. However, the ambiguity between the query terms and the different senses of the words in the WordNet has been the most challenging problem in using WordNet in query expansion. Here, we combined two methodologies, described below, to capture the context information of the words entered by the user to reduce the ambiguity.

2.1.1 Selecting sense from Synsnet

In linguistics, a word "sense" is one of the meanings of a word in dictionary. Similarly, finding the proper sense of a word in WordNet is one of the major steps in determining the semantics of the query terms and reducing the disambiguity while extending the words [27,35,42]. For example, the word "Java" has several senses. It can refer to an island in Indonesia, coffee or a programming language based on the sentence it is used. Thus, there is a need for context information to identify the correct sense of the word. The set of words entered in the query can form a context to use for disambiguating the ambiguous word. In this case, if the set of words entered with java are "bean," "mug" and "bold," we will find out that the second sense is the most appropriate sense describing the meaning of "Java" in a given context (Fig. 2). [33] introduced WordNet::SenseRelate::WordToSet Perl package in which the context of a

```
C:\strawberry\perl\sire\bin>perl wordtoset.pl –type WordNet::Similarity::jcn java land region population

java#n#1 : 0.32837470147 : an island in Indonesia south of Borneo; one of the world's most densely populated regions
java#n#2 : 0.22974195638 : a beverage consisting of an infusion of ground coffee beans; "he ordered a cup of coffee"

C:\strawberry\perl\sire\bin>perl wordtoset.pl –type WordNet::Similarity::jcn java bean mug bold

java#n#2 : 0.18195734024 : a beverage consisting of an infusion of ground coffee beans; "he ordered a cup of coffee"
java#n#1 : 0.14046287135 : an island in Indonesia south of Borneo; one of the world's most densely populated regions
```

Fig. 2 Sense disambiguation of the word java based on the context words

word is defined as the associated set of words. Thus, the correct sense of a target word is projected to be the sense that is most related to the words in the associated set. If there are N context words in the associated set of words, the best sense of the target word w is calculated as:

$$Best\_Sense = \arg\max_i \sum_{j=1}^{N} \max_k relatedness(t_i, s_{jk})$$

where t_i sense of the target is word, and s_{jk} is sense k of the jth context word.

Following the definition, in this research we use WordNet::SenseRelate::WordToSet Perl package with WordNet version 3.0 to find the sense of the words mapped as a result of "Phrasal Map." The SenseRelate package uses different relatedness measures for the similarity calculation [37]. In the scope of this research, all the mappings and comparisons are based on the WordNet noun set. Thus, as a robust relatedness measure for a noun-noun similarity [38], the jcn [24] content-based similarity measure is used to determine the relatedness of the words entered in the query. Consequently, the highest ranked sense of the word is used to choose the correct sense of the words from WordNet (Fig. 2).

The aim of finding the correct sense of a word is to extract the semantically related terms that potentially improve the relevancy of the returned documents for user's request. To achieve this, in this research we have used "hyponym," "holonym" and "synonyms" of the mapped words for the corresponding sense that is found as a result of SenseRelate algorithm.

2.1.2 Phrasal map

Once user input is received, CONIA uses stop-word elimination to peal improper words in the documents. The remaining words are used to pass to WordNet for evaluation. It is important to note that word-by-word mapping between the query terms and WordNet does not always preserve the context of the terms provided in the query. For example when "brain tumor" is passed to WordNet, "glioma" is returned as its subclass. However when "brain" and "tumor" are passed to WordNet separately, hundred of subclasses of tumor are returned including such as "lymphoma" and "Hodgkin's disease," which are not related to the "brain tumor."

In order to better predict the context of the user's input, we find all the combinations of the query terms before mapping them into the WordNet. Given a set of n terms in the query, all combinations of the size r, $C(n, r)$, are calculated and added to the set W as word/phrase. Starting from the set with max number of terms, each set is checked against WordNet for possible match. If a match is found, all the other combinations in W that are subsets of the matched phrase are erased from the list (see Algorithm: Mapping from Query to WordNet). For the ones that are mapped (query-mapped terms), their synonyms are added to the combination set in which they are replaced by the mapped words for derived combinations. The example of the query and mapping is presented as follows:

This mapping makes it possible to better estimate the context of the query words by allowing more than one term to be mapped from query to WordNet. Multiple word mappings usually correspond to more specific concepts in WordNet than the single terms that help to reduce the ambiguity of the words by linking them together.

Algorithm: Mapping from Query to WordNet

Return 1: Set of mapped words M

Return 2: Set of derived combinations including synonyms WS

```
Q: the set of words in the query after the stop word elimination
W: set holds all possible combinations.
S: set of synonyms of Mapped words
WS: Mapped words and derived synonyms.
M: mapped words in WordNet
n: Number of words in Q
r: An integer, 0<r<=n
r=n;
While (r! =0) {
   Find Combinations of n in size r (Combination (n, r))
   Add combinations to W
   r=r−1;
 }
For (i=n; i>0; i–) {
   For (j=n; j>0; j–) {
     If wi is not subset of mi in M {
     k= Subset;
   }
   If k ≠ Subset{
Map wi to WordNet ()
If wi is in the WordNet {
        Add wi to M
}}}}
WS=M;
For each mi in M {
     If Synonym si of mi{
   For each mj contains mi{
        Replace mi with si
          Add mj into WS
     }}}
```

2.2 Customizing user's information request

Upon completing the word mapping, customization is used to get relevant results. Customizing the results returned by Information Retrieval Systems for each user is a challenging task. Many research studies have been done on relevance feedback, to collect the data about the user behavior [2]. However, all these research is affected by the drawback of human interaction. In this paper, we are going to take the advantage of the common domain interest of the users to predict their information needs without the need of their interaction. Although it is not a restriction, the main use of CONIA is to find the information need of the users working on a specific domain. This assumption gives us an opportunity to reduce the ambiguity of the words that can be derived from the query in addition of focusing the common goals of the domain users.

In this paper, we chose "Health" and "Terrorism" domains as example domain to integrate with CONIA. The returned set of documents for the information request of users in these

domains can be used for critical decisions. Time and information accuracy are also critical components of the domains. The users of these domains are in need of harvesting most up-to-date, reliable information in the shortest available time, which is a challenge problem for Information Retrieval techniques.

2.2.1 Interest-ontology

Although, the domain ontology provides information about the domain-specific aspects, when it comes to user request, it becomes clear that the users within the domain might have different interest. For example, while one health professional is looking for information about "cancer" in children, the other might be interested in finding information about the "cancer" in elderly people. Thus, the context of a user's request within the domain can also be affected by user's interest. Therefore, user's interest within the domain named as "interest-ontology" and defined as follows:

> A small ontology with general concepts belonging to a specific domain that is different than the domain ontology but relates to it, is called Interest-ontology. In other words, it is a high level domain ontology that shares common concepts or relations with the domain ontology used by the system.

In the following section, we are going to give the details of example "Health" ontology to demonstrate the idea discussed here.

2.2.2 Domain ontologies

Fodeh et al. states that the ontology can be used to greatly reduce the number of features needed to do document clustering. One of the challenges of using domain ontology is finding a well-established, detailed ontology about the domain. The "Health" and "Terrorism" ontologies that are going to be explained in this section are implemented by using Web Ontology Language (OWL), and Protégé 3.3 is used for visualization and manipulation of this ontology.

2.2.2.1 Health domain ontology The National Cancer Institute's (NCI) Thesaurus is an ontology-like vocabulary that includes broad coverage of the cancer domain, including cancer-related diseases, findings and abnormalities; anatomy; genes and gene products, drugs and chemicals; and organisms. In certain areas, like cancer diseases and combination chemotherapies, it provides the most granular and consistent terminology available. It also combines terminology from numerous cancer research-related domains and provides a way to integrate or link these kinds of information together through semantic relationships. Thesaurus currently contains over 34,000 concepts, structured into 20 taxonomic trees. In the scope of our research, we have eliminated some branches of the NCI ontology such as organisms, gene and gene products and drop the size of concepts to 15,258 to make is easier to run in a laptop with 2 GB RAM. The extracted NCI ontology is a very shallow with max depth of 5, max sibling number of 814 and very few number of associative relationships (Fig. 3).

As we have mentioned in the previous section, users of the domain might have different interest in their information request. To demonstrate this idea, we built 2 interest ontologies for the health domain: "Environment" and "Person." The Environment interest-ontology covers the concepts from the environment that cause health problems. For example, air borne pollutants connect with the respiratory problems in the health domain or commuting in traffic

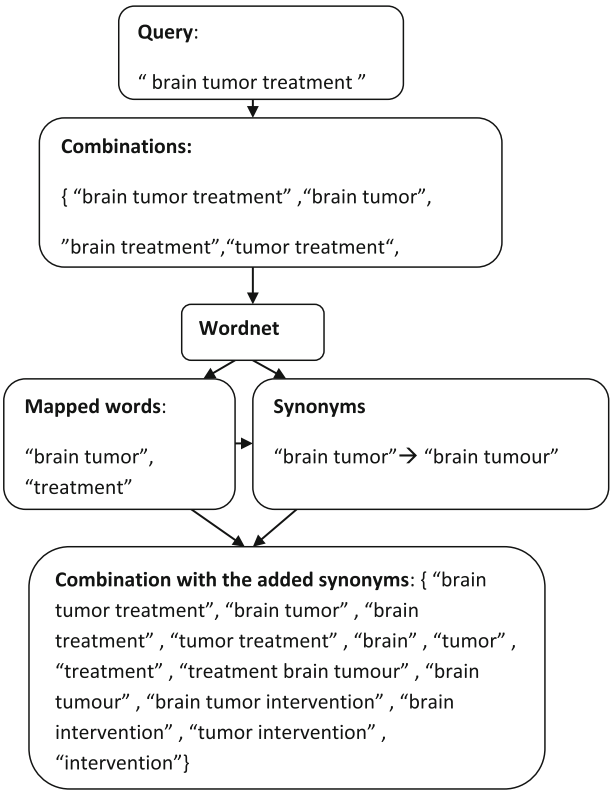

Fig. 3 Pharsal map example

causes stress and dependently depression problem (Fig. 4a). The concepts of "Environment" interest-ontology is taken from US Environmental Protection Agency [50] and organized in hierarchy. "Person" interest-ontology is used to characterize the people with the health problems. For example, the user searching for leukemia might be interested with the information sources that talk about leukemia in the context of children.

Ontology integration is an important research area for ontologies [5,9]. Having an inconsistencies is almost inavoidable when more than one ontologies are integrated. Even if two systems adopt the same vocabulary, there is no guarantee that they can agree on a certain information unless they commit to the same conceptualization [16]. Assuming that each system has its own conceptualization, a necessary condition in order to make an agreement possible is even harder.

Since the aim of this research is not ontology integration, in order to ease the communication between the interest ontologies and domain ontology, starting from the root node we added each interest-ontology as a different branch in the domain ontology (Fig. 4b). Then, the relations between the interest-ontology branch and the health domain branches are provided by the associative relations.

2.2.2.2 Terrorism domain ontology Since September 11, 2001, the focus of mining terrorism data to predict further attacks by extracting patterns becomes an interest of Intelligence

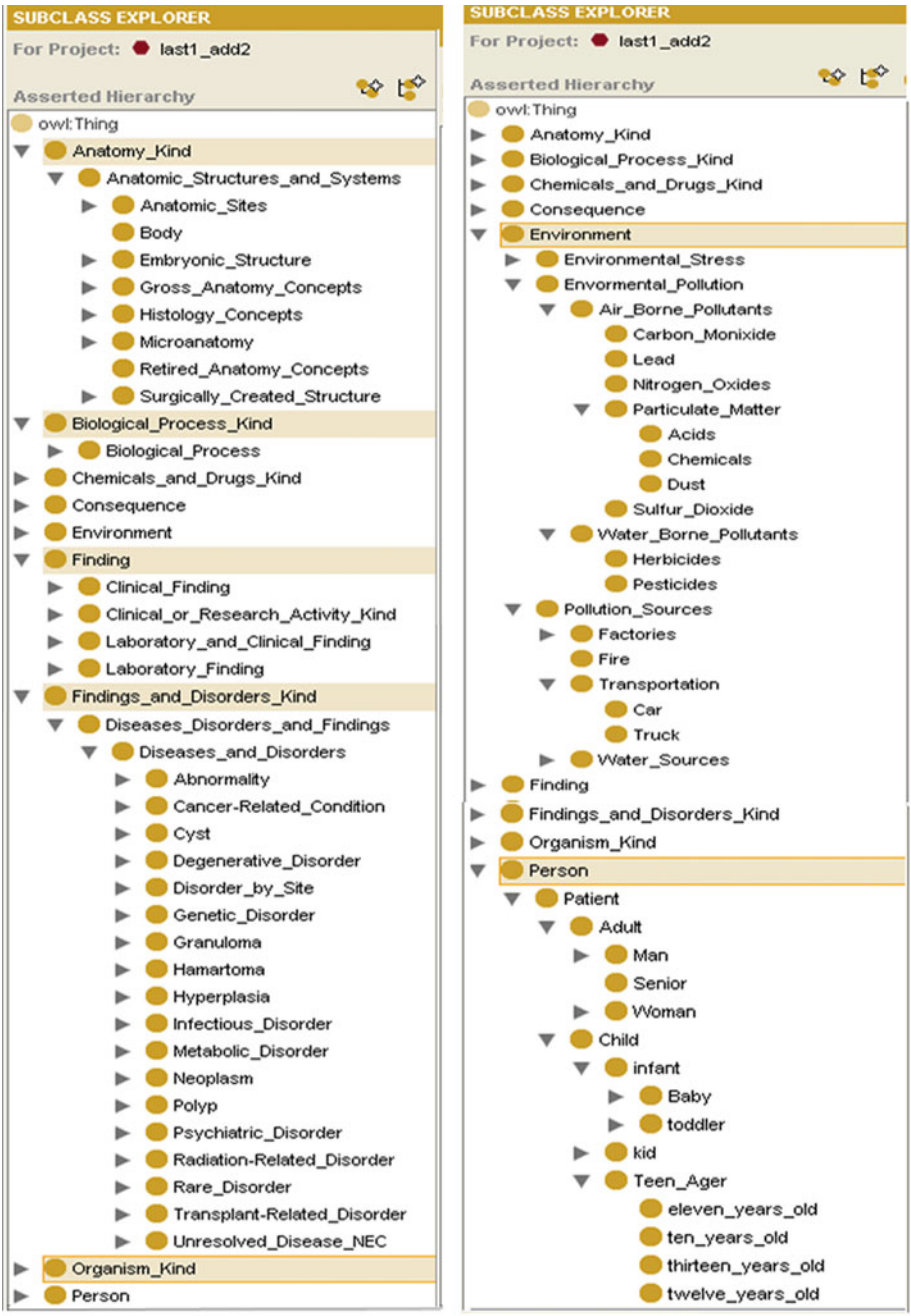

Fig. 4 **a** Customized NCI ontology and **b** interest ontologies integrated into NCI ontology

Analysts. The need of finding up-to-date data about terrorism events suggested that the terrorism domain should be one of the example domains of this paper. However, the top secret category of the domain makes it difficult to find a publicly available "Terrorism" ontology.

Between September 14, 2001 and November, 2001, Krebs assembled a corpus of texts regarding events preceding September 11 attacks. In the aftermath of the September 11 attacks, it was noted that coherent information sources on terrorism and terrorist groups were not available to researchers [15]. Information was either available in fragmentary form, not allowing comparison studies across incidents, groups or tactics, or made available in written articles, which are not readily suitable for quantitative analysis of terrorist networks.

To counter the information scarcity, a number of institutions developed unified database services that collected and made available publicly accessible information on terrorist organizations. This information is largely collected from open-source media. Such open-source databases include the following: RAND Terrorism Chronology Database [40], including international terror incidents between 1968 and 1997, RAND-MIPT (Memorial Institute for Prevention of Terrorism), Terrorism Incident Database [22] and MIPT Indictment Database [47]—Terrorist indictments in the United States since 1978.

Both RAND and MIPT databases rely on publicly available information from reputable information sources, such as newspapers, radio and television. Gruenwald et al. [15] introduced architecture for extracting ontology from the available databases. Military of Defense (2005) used small "Terrorism" ontology in a case study. Defense Research and Development in Canada introduced Auger [1] as an Information Retrieval technology for "Terrorism" domain in which it retrieved the information about tactics, weapons, targets, persons, groups and locations. In [32], Mannes and Golbeck proposed Mindswap terrorism ontology that is constructed by the cooperation with terrorism specialists.

Despite these few attempts on constructing terrorism ontology, we only had an access to the Mindswap terrorism ontology and found excerpts from MoD and TerroGate ontologies. Based on our research interest on the topic for years, we have constructed the "Terrorism" ontology used in this research from these three ontologies. The "Terrorism" ontology we built consists of six major branches: Event, Location, Organization, Person, Target and Weapons (Fig. 5a). These are the common categories used in the databases and the ontologies constructed by the other researchers mentioned above. The "Terrorism" ontology consists of 128 concepts with max dept of 4 and max number of siblings 12. Although it is smaller than the "Health" domain ontology, the associative relationships in this ontology is denser. For example, some of the terrorist groups are connected with events based on the statistical information provided about the terrorist groups, and the activities are represented in the terrorism ontology [18].

Two interest ontologies, "Economics" and "Politics," are constructed by using the information from MUC4 terrorism news [36] in addition to the other articles in the internet (Fig. 5b). The "Economics" interest-ontology consists of the concept that relates to the effects of a terrorist attack on economy, while the "Politics" interest-ontology includes concepts that relate the terrorist attacks to politics.

3 Mapping form WordNet to domain ontology

As explained in Sect. 2.1.2, after mapping query terms to WordNet, the following sets are obtained: set of mapped query terms from query to WordNet, is-a and part-of subclasses of mapped query terms, and all the possible combinations of the query terms including the synonyms. Although the contextual relations of the query terms are used in the sense selection of the terms, the information about the domain is still missing.

In order to better eliminate the ambiguity of the obtained set of words from WordNet, there is a need of selecting a subset of the expanded set of words, which are relevant in the context

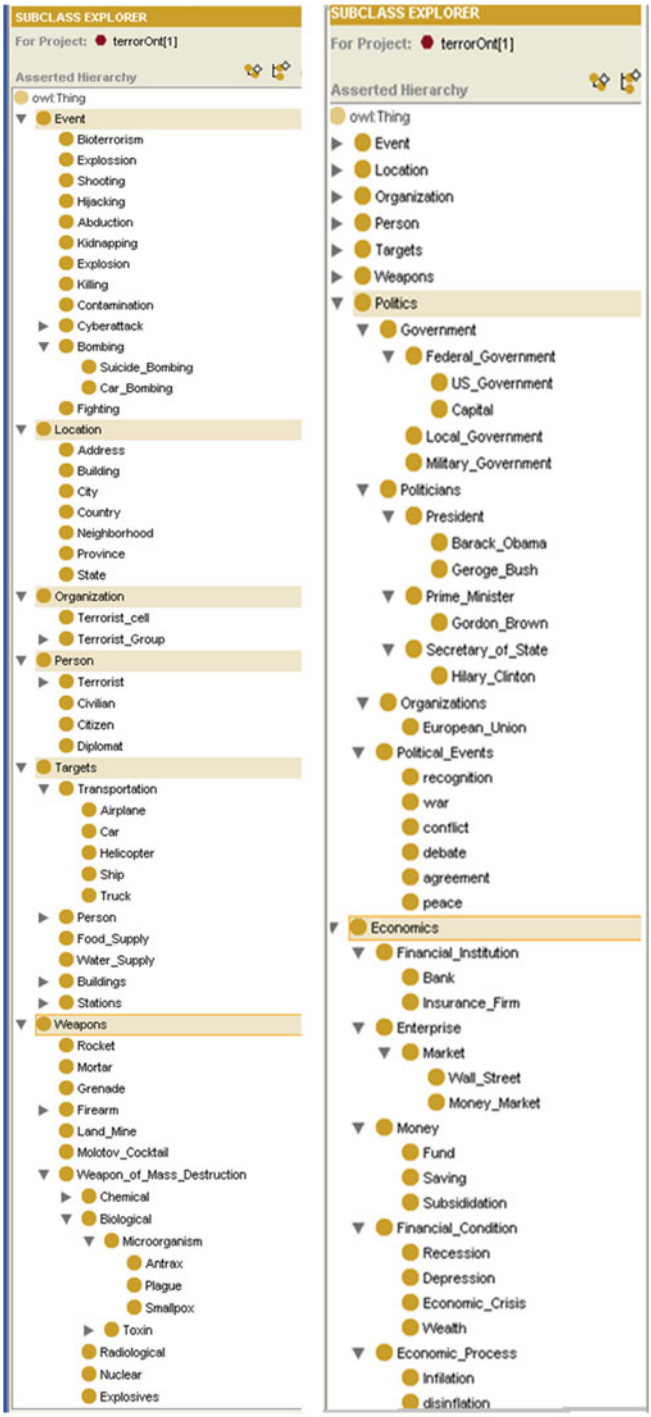

Fig. 5 **a** Terrorism ontology and **b** integrated interest and terrorism ontologies

of the domain user's request. Before explaining the mapping process between WordNet and the domain ontology, we would like to define the word "Ontology Category" that will be used in the rest of the paper:

> Each branch starting from the node that has the root as a first level ancestor is called an ontology category

For example in the "Health" domain ontology (Fig. 4a), the root is "Thing;" thus, "Anatomy Kind," "Biological Process Kind," "Chemicals and Drugs Kind," "Consequence," "Environment," "Finding," "Finding and Disorders Kind," "Organism Kind" and "Person" are the ontology categories. The reason of defining categories within the ontology comes from the fact that each branch of ontology covers different kind of information. More information gathered about the different aspects on the ontology, the better view of the query context can be observed.

3.1 Syntactic and semantic map from WordNet

The mapped set of terms obtained from WordNet in Pharsal Map (e.g., "brain tumor" and "diagnosis") are based on syntactic map from query to WordNet. Although the context of query terms is used to extract the proper sense of the words, it is not guaranteed that the terms entered in the query will have any coherence for the determination of relevant senses of terms.

Mapping the terms obtained from query to the concepts of the domain ontology is a challenging process [31]. In oppose to mapping ontologies, there isn't as much research on mapping the query terms to the ontologies. In some cases, ontology query languages are used instead of free style request [54].

Different than the lexical ontology such as WordNet, in which concepts are composed of mostly single or two, three words that together have a definition, domain ontology concepts can be composed of subsentences in which might not have a definition or synonym (e.g., "Generation of Antibody Diversity" and "Monoclonal Antibody M170"). Mapping the set of words obtained from WordNet to an ontology that consists of unstructured concepts is a challenging task. A word such as "antibody" can occur in different ontology categories and might be part of a concept that consists of multiple words. Therefore, there is a need of a decision mechanism to decide which mapped concepts best matches to the corresponding query terms. The following steps are applied in an order to make this decision:

- Map each query-mapped term (qmt), its subclasses and the synonyms of qmt, found as a result of phrasal phrase, from WordNet to the domain ontology.
- If two subclasses (helonym) of qmt has one-to-one map (qmt-mapped subclass) to the concepts in the domain ontology and they are siblings in the proximity of 2 or less (have a common ancestor in proximity of 2 or less), then keep those subclasses. The depth 2 is taken since the ontologies used in this research are not too deep and mostly shallow. For example in (Fig. 6a), "Truck" and "Hospital" have "Target" ancestor in proximity 2.
- If one qmt-mapped subclass is an ancestor of the other qmt-mapped subclass in the domain ontology by the proximity of 2 or less, then keep these subclasses (Fig. 6b).
- If any of the mapped qmt on the domain ontology is an ancestor of any other qmt-mapped subclass in the proximity of 2 or less, then keep the subclass (Fig. 6c).
- Categorize picked subclasses based on the ontology categories and find the number of subclasses in each category.

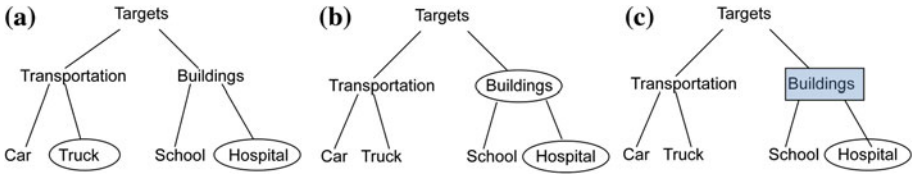

Fig. 6 **a** Two sibling qmt-mapped subclass, **b** one qmt-mapped subclass is ancestor of the other, **c** qmt-mapped term and qmt-mapped subclass

Mapping qmt subclasses to the domain ontology and extracting the common subclasses through the above cases help us to find the concepts of the domain ontology that are semantically related to the query. However, besides the semantic matching of qmt subclasses, the importance of the exact words entered by the user should be considered in assigning the relevant Ontology Categories on the domain ontologies. The selection processes of the most relevant Ontology Category of the domain ontologies without associative relations are done as follows:

- Count the number of qmt maps in each Ontology Category.
- If there is an exact match (one-to-one mapping) between qmt and the domain ontology concepts, then assign qmt to that ontology category.
- If the qmt doesn't have an exact match but all the maps are in the single ontology category, then assign qmt to that ontology category.

When mapped qmt concepts correspond to more than one Ontology Category, there is a need to choose the proper concept and the belonging Ontology Category. For the ontologies that qmt maps to more than one category and do not have an associative relationship between the mapped qmt concepts, the following criteria are used to choose the best category qmt belongs.

- If qmt maps more than one category, then check the category of the selected qmt subclass and assign qmt to the category of qmt subclass.
- If there is more than one mapped categories and the number of mapped cases between the two highest categories is less than 5, then check whether any other qmt is assigned to the category with highest number by the above-mentioned steps, if so pick the second highest category. Otherwise, pick the category with the highest number.
- If there is a query term that is not part of any qmt or has no maps in domain ontology, then use synonyms as with all the combinations derived from query terms to check if there is any map on domain ontology. If so, apply the above steps.
- If neither synonyms, nor combinations with synonyms of a query term appear in the domain ontology, then assign it to the category called "Independent" in which means it is not available in the ontology.

Once the above process is completed, each term/terms entered by the user in the query is assigned to a category in domain ontology with the subclasses relevant to the domain. As a result, we have the subset of the extracted words from WordNet. However, domain ontology is more likely to have detailed information about the mapped concepts. Thus, once the mapped concepts are found, their extracted subclasses from the domain ontology are added to the set of extended terms.

The above-mentioned algorithm is good when there isn't any associative relationship between the ontology categories. But the real gain of using ontologies is to extract the

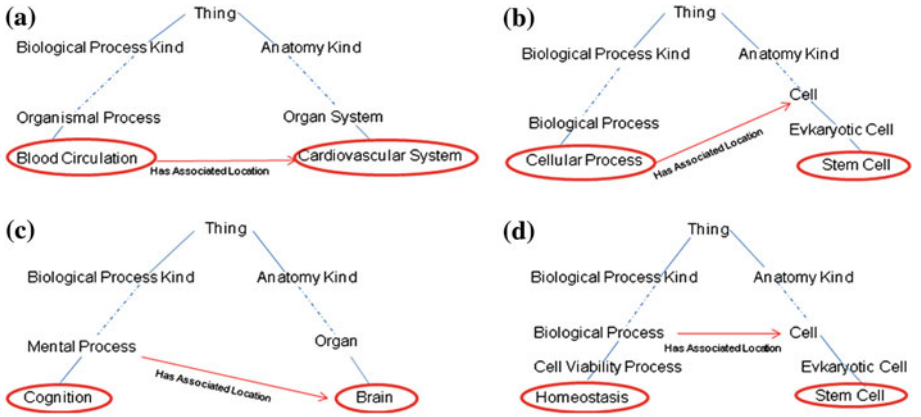

Fig. 7 **a** Direct associative relation, **b** associative relation to a parent, **c** associative relation from a parent, **d** associative relation between the parents

semantic information provided through associative relations. Therefore, the following cases are used to select the proper categories of mapped qmt concepts in the availability of associative relations between the mapped qmt on the domain ontology. In Fig. 7, qmt-mapped concepts of NCI ontology are represented in red circle, and their categories are assigned as "Biological Process Kind" and "Anatomy Kind."

Case I: If there is a direct associative relation from one qmt-mapped concept to the other one that is belonging to different ontology category, then choose two qmt terms and their categories (Fig. 7a).
Case II: If there is a relation between one of the mapped qmt concepts and the parent of another mapped qmt concept at the proximity of 2 or less that belongs to a different Ontology Category, then choose the mapped qmt concepts, categories with the connecting concepts and their relations (Fig. 7b).
Case III: If there is a relation between the parent of a mapped qmt concept and the other mapped qmt concept at the proximity of 2 or less that belongs to a different Ontology Category, then pick mapped qmt concepts, categories with the connecting concepts and their relations (Fig. 7c)
Case IV: If there isn't any direct relation between the mapped qmt concepts but their ancestors are in the proximity of 2 or less, then pick the mapped qmt concepts, categories with the connecting concepts and their relations (Fig. 7d).

3.1.1 Finding relations between qmt in domain ontology

As a result of the above cases, we observe the paths using associative relations between the mapped qmt concepts. Here, the terms "Left Hand Side" (LHS) and "Right Hand Side" (RHS) are used to explain the associative relations within the paths. Anything before the associative relation on the path is considered as the LHS, whereas anything after the associative relation is considered as RHS. For example in Fig. 7d, the path is "Homeostasis, Cell Viability Process, Biological Process → Cell, Eukaryotic Cell and Stem Cell." In this path, "Homeostasis, Cell Viability Process and Biological Process" is LHS of the path, while "Cell, Eukaryotic Cell and Stem Cell" is the RHS of the path. In case there is more than one associative relation in the path, everything before the last associative relation is considered

as LHS and after the last associative relation is considered as RHS of a path. The following criteria are used to extend the paths such as P1 and P2 that share common concept.

- If the first concept say "X" on the RHS of P1 appears on the LHS of P2:
  - Take all the concepts and the relations of P1 before "X" and combine with the all concepts and the relation that appears after "X" in P2. For example, assuming that P1: A,B,C → X,E and P2: F,X,V → R,T, we can derive a new path P: A,B,C → X,V → R,T
- If a concept say "X" that is in the LHS of P1 appears on the LHS of P2:
  - Take all the concepts on left of concept "X" from P1 and combine with all the concept and relations that is on the right of concept "X" from P2 and vice versa. For example, assuming that P1: A,B,X,D → E,F and P2: T,K,X,L→P,R, we can derive P:A,B,X,L→P,R and P′:T, K, X, D →E,F
- If a concept say "X" that is in the RHS of P1 appears on the RHS of P2:
  - Take all the concepts and relations before "X" from P1 and combine with all the concept and relations that is on the right of concept "X" from P2 and vice versa. For example, assuming P1: A,B,C → D,X,F and P2: E,G →T,X,H, we can derive P: A,B,C → D,X,H and P′: E,G → T,X,F

The above cases provide all the possible paths between mapped qmt on the ontology. Since these paths provide us coherent information about the overall context of the query words, it is important to choose the mapped qmt and the assigned categories that are belonging to these paths. Longer the path, more likely to have more mapped qmt and associative relations that cross between different ontology categories. Once the proper paths (the one with most mapped qmt that belongs to different ontology categories) are selected, subclasses of the selected concepts in these paths are also added to the set of chosen terms to extend the word list.

3.2 Ontology categories and context

In oppose to "Static" Informational Search in which user usually enters a single word to find the information about description/definition he requests, the users with the intention of finding "Dynamic" Information tend to enter more than one terms to define the context of a search. Passing the query terms to WordNet mostly helps to observe the related terms corresponding to the same concept category. Once terms mapped from WordNet to the domain ontology, the information about the terms that are relevant to extend in the context of a domain is obtained with their assigned ontology categories.

One of the main goals of our research is to show that successful IR should take the context information in bringing the relevant documents. In this paper, ontology categories are the representative of the different context elements within the domain. Thus, the *context of the query terms* within the domain is defined as follows:

> Query terms mapped to the domain ontology under different ontology categories construct different context elements of the query.

Context is mostly associated with the events. Context-based event extraction [3] is a research area by itself which is not the main focus of this research; however, analyzing the event information within the domain well suits to the goal of "Dynamic" Information search. Thus, we define the *event category* within the domain ontology as follows:

An ontology category can be called as an event category iff:

- Concepts of the ontology category has a consequence
- Concepts of the category is a consequence of the concepts from another ontology category (post_ event)

In the domains like "Terrorism" and "Health," the focus of the domain events is people; thus, the consequence of event category expected to affect people. For example in the health domain, the category "finding and disorder kind" can be assigned as an event category of the domain since the disease affects people and as a consequence people get sick, die, get blind and so on. At the same time, the ontology category "Consequence" is also can be defined as an event category within the domain since it is an event that occurs as a consequence of the other event. In mapping the extended terms to the documents to find which document potentially provides the best relevant information for the user's request, the context information of the mapped terms is considered.

3.3 Document set

The document set used as a corpus in this research is collected through Web. CONIA is not a crawler but instead uses other search engines to collect the data for its corpus. When user enters information request to CONIA, query passes to two modules. One module is WordNet and ontologies as explained in the previous sections, and the other is the search engine. Yahoo is the main search engine used to collect the documents for the corpus of CONIA. Since the user's of CONIA is interested in finding the most up-to-date dynamic information, primarily Yahoo-news search is used to collect the relevant documents.

In addition to the advantage of using other search engines to retrieve the documents, not having a control on crawling process comes with some disadvantages. The selection of documents in the corpus of CONIA depends on the relevance criteria used by the search engine. This brings limitation to CONIA since in some cases, the corpus might not have any documents matching to CONIA's relevancy criteria even there are relevant document corresponding to the user's request in the Web.

Similarly, the Web environment itself has some pluses and minuses. While CONIA takes the advantage of the dynamic structure, domain free data set, easy access and no pre-processions; huge number of documents, repeated documents, formatting and trust issues are among the challenging issues.

In the following sections, we are going to explain how we incorporate the above-mentioned challenges within the structure of CONIA.

3.4 Ontology to document map

In the previous sections, we have explained how the terms inputted in the information request of a user is expanded and also explained how the documents used as a corpus of CONIA is constructed. In this section, the process of mapping expanded query terms to the documents are going to be described.

3.4.1 Capturing concepts in the document

The traditional IR techniques based on indexing uses number of mapped words appear in the document to decide on the relevancy of the documents. However, this technique does not

consider any semantics or the context information that is crucial in user's information need. The motive behind CONIA is to add the domain and context information into document selection process and customize returned documents based on users' interest.

The main reason of assigning concepts into Ontology Categories as mentioned in the previous section is to use these categories as context elements in the process of mapping expanded words to the documents. Once all the words extracted from domain ontology are mapped to the document, there is a need to analyze how the mapped words contextually fit within the document. In this paper, the contextual relation of the mapped words within the document is based on the actual distance between different category elements.

The distance can be computed in different magnitudes. WordNet defines "Paragraph" as "*one of several distinct subdivisions of a text intended to separate ideas."* Since our interest is to find the semantically relevant documents for user's information request, we used paragraphs as a measure to calculate the distance between the elements of different ontology categories.

3.5 Finding content relevant documents

In order to retrieve the relevant documents in response to user's request, a value should be assigned to the documents. In this paper, the relevancy value of the document is assigned in paragraph level. First, the extended set of words with their context categories (ontology categories) are found in the paragraph, then the minimal distance between the different contexts elements is calculated as will be explained in the following sections. Once the minimum distance between the mapped words is calculated, the value of the paragraph is compared with the value of all the other paragraphs and the paragraph with the minimal contextual distance selected to represent the relevance value of the document.

Algorithm: Calculating Content Relevancy
Return: Relevance Value of the Document

```
P: Set of Paragraphs in the document
D : Document
W: Set of all expanded words
M: List of words mapped from ontology to the Paragraph
MC: Category of each Mapped Word

MinDist=100000;
For each (P: p1, p2…, pk) {
    M← MappedWords (W, pi)
    For each (M: m1, m2…, mn) {
        MC← FindCategory(mj)
    }

Distance ←OverallMinDistanceBetween DifferentCategoryElements (MC)
    If Distance < MinDist;
        MinDist=Distance
    }
    Document Relevance= 1/MinDistance;
```

3.5.1 Proximity of mapped terms

Term proximity has been explored extensively in document ranking studies [8,19,41], where several distance factors were proposed. Two common intuitions underlie in all of these approaches:

- The closer the terms are in a document, the more likely they are topically related.
- The closer the query terms are in a document, the more likely the document is relevant to the query.

Our intuition in this research is also same. The difference is, in the above methods, term proximity is calculated in the document level, and all the terms are treated equally. In CONIA, terms are further customized by the following criteria:

- Terms are associated with context categories and the proximity of terms is calculated between different context category terms.
- Term Proximity of the document is calculated in the paragraph level.

Thus, we adopted two of the following approaches from the above studies.

3.5.1.1 Span-based proximity measure *Span* is defined as the length of the shortest document segment that covers all query term occurrences in a document including repeated occurrences [19].

Similarly, *MinCover* is defined as the length of the shortest document segment that covers each query term at least once in a document.

The term proximity in this paper is adopted from MinCover and called as CMinSpan. Given a paragraph, CMinSpan defined as the length of the shortest document segment within the paragraph that covers at least one term from the each context category that appears in the paragraph. For example, for a given paragraph p, let t_i, a_i, b_i and c_i represent the terms. If a_i represent the terms under category 1, b_i represent the terms under category 2, c_i represent the terms under category 3 and t_i represent the terms that do not belong to any categories in the domain, for the given paragraph $p = (t_1, t_2, a_1, t_4, c_2, b_1, t_5, t_6, a_2, t_1, t_2, c_1)$, CMinSpan value calculated between the shortest document segment that covers 3 categories which is $\{a_1, c_2, b_1\}$, and the score for this calculation will be assigned as 4 that is the length of the word span within this interval.

Both Span and MinCover methods favor documents with fewer query occurrences. Normalization factor is used to fix the bios by dividing Span value with number of occurrences of query terms in span segment [49]. In CMinSpan normalization is done based on the number of categories that will be explained in the Sect. 3.5.1.2

3.5.1.2 Pair-wise proximity measure *Pair-wise* distance is defined as a distance between individual term occurrences, and the overall proximity distance value is calculated by aggregating pair-wise distances. For example for the query words $q = \{t_1, t_2, t_3\}$, the proximity between the term pairs $\{(t_1, t_2), (t_1, t_3), (t_2, t_3)\}$ are calculated. Rasolfo et al. [41] compute term pair instance weight as follows:

$$tpi(t_i, t_j) = \frac{1}{d(t_i, t_j)^2} \tag{1}$$

where $d(t_i, t_j)$ is the distance expressed in number of words between search term ti and tj.

We adapted Rasolfo's *tpi* formula in the scope of CONIA. However, instead of calculating the distance between each mapped term in the document, the distance between each mapped

terms of a paragraph that belongs to a different category is calculated. For the example given in CMinSpan, $p = (t_1, t_2, a_1, t_4, c_2, b_1, t_5, t_6, a_2, t_1, t_2, c_1)$, CMinSpan $\{a_1, xc_2, b_1\}$, pair-wise distance would be calculated between $\{a_1, c_2\}, \{a_1, b_1\}, \{c_2, b_1\}$. Thus, we define term pair weighting between two terms of different categories as:

$$Tpw(c_i, c_j) = \frac{1}{\min(d(c_i, c_j))} \tag{2}$$

where $d(c_i, c_j)$ is the distance expressed in number of words between the two mapped terms (formula (1)) in the paragraph that belongs to different categories. If a category has more than one terms mapped, then the minimum distance is selected to represent $d(c_i, c_j)$. In (formula (2)), we eliminated the square of the distance since the distance of the terms in our approach is calculated in paragraph level other than the document and won't be that big.

We want to go further and customize (formula (2)) more, to fit the purpose of the applications used in this research. As we mentioned in the previous sections, it would be logical to choose an event category among the Ontology Categories. A user searching about event-related information is most probably to be interested with the context elements of the event (in which will be the associative relations from event category in domain ontology). Thus for the paragraphs with more than two different category terms mapped, first we are interested in finding whether there is a mapped term to the event category and if so we will find the minimum pair-wise distance from the term belonging to the event category to all the other categories. For instance, in the above example if "c_2" would be the term from the assigned event category, the pair-wise distance between the terms $\{c_2, a_l\}, \{c_2, b_1\}$ would be calculated. Therefore, the total pair-wise term proximity of the paragraph with mapped event term is calculated as:

$$TTpw(E, c_j) = \frac{1}{\sum_{c_j \in C} Tpw(E, c_j)} \tag{3}$$

where $\sum_{c_j \in C} Tpw(E, c_j)$ is the sum of the minimum distances from a term E that is a member of event category, to a term from the other mapped category in the paragraph. If the paragraph has more than one term mapped from event category, in this case the minimum of *TTpw* is selected

$$MTTpw(p) = \min(TTpw(E, c_j)) \tag{4}$$

In the following section, we are going to explain how CMinSpan and Pair-wise distance methods are used and normalized in the scope of CONIA.

3.5.2 Content relevance value of the document

The content-based relevance of a document is calculated by the term proximity values obtained from its paragraphs. The paragraph level term proximity is calculated by the methods introduced (CMinSpan, Pair-wise). If any paragraph of the document has a term from the event category, then both CMinSpan and Pair-wise methods are used in ranking the documents. Otherwise, if there are no paragraphs with a term from event category, CMinSpan is used to calculate the value of a document.

Other than the terms in the event category, the ranking of document is affected by the number of different category elements available in the paragraphs. The number of distinct categories mapped on the paragraph is the main measure on determining the relevance of the documents. More the different category terms mapped to the paragraph, the paragraph

is more likely to have a coherent theme that corresponds to the user's request. If interest-ontology is selected along with the request, the interest-ontology is treated as a part of the domain ontology in the term mapping. Besides, in order to better determine the focus of the document in a particular interest, the number of distinct terms within the document and the paragraph level that fall into the interest category is calculated. Overall document relevancy is calculated in the following order:

- The paragraph with the highest number of mapped categories is selected to represent the relevance value of the document. Each document is categorized into levels based on the number of different categories of the selected paragraph.
- Depending on the category kind, CMinSpan and/or Pair-wise proximity measure is used to calculate the proximity value of the paragraph with the highest number of distinct categories.
- If more than one paragraph has highest number of categories, then the paragraph with the max proximity value is chosen (max = 1/min).
- In case user picks "interest-ontology" while requesting the information, the number of distinct terms within the paragraph and the document was calculated to determine the emphasis of the document on the interest area of the user.

The most of the Span-based proximity calculation methods favor on the documents that have few words mapped and need normalization to balance the proximity value. Usually the number of words within the document is used for normalizing the proximity value [45]. In this research, number of categories and the document with the max number of relevance is used for normalizing CMinSpan value as follows:

- The document with the highest proximity value gets selected, and the proximity value of all the other documents is divided by that value.
- Documents are sorted within each level based on the proximity value.

Finally, documents are sorted from most relevant to the least relevant for the user's request.

3.5.3 Content extraction from HTML documents

As mentioned in the previous sections, the documents returned by the search engine might be in different formats. The documents used in this research are in HTML format. In order to analyze the paragraph structure of the documents, we pass them through an open-source html parser (http://htmlparser.sourceforge.net/). HTML Parser is a Java library used to parse HTML in either a linear or nested fashion. The parser produces stream of tag objects, which can be further parsed into a searchable tree structure.

Our analysis on the retrieved news documents showed that most of the documents use paragraph "<p>" tag to capture the paragraphs within the body of the documents. In addition, in some cases we also observed that "

" and "</br></br>" tags are used to set a new line between the paragraphs. Since </br> tag is not a valid tag by the html parser, we have used "<p>" and "

" tags to parse the html documents that cover about 90 % of the documents retrieved by the search engine.

4 Experiments

The aim of this section is to present the experimental results for CONIA and the other IR systems to measure the success of the proposed system. One of the most challenging tasks

in IR is to measure the success of a system. The major challenge is finding the data set that can be used by several IR for comparison. Since the relevance of a document is relative to the user, the way of measuring success of any system is based on the users' feedback.

The data set used by CONIA is dynamic and crawled by currently available search engines (e.g., Google/Yahoo). Therefore, there is no pre-evaluated set of documents to compare the results of CONIA. In an open environment in which data set is dynamically changing, the success of the systems is often measured by the user evaluation [14,25]. For example, Liu [28] used 25 graduate Computer Science students for 3 weeks to measure the success of Google, Yahoo and MSN search engines. Similarly, Long et al. [30] recruited 270 participants, who were from ten universities of Beijing to test the three Chinese search engines.

Taking the above methods as an example, in this research we used human rating to evaluate the success of CONIA.

4.1 Setup of the experiment

Most of the IR systems are compared in two aspects: precision and recall. Unlike having a corpus of pre-collected documents, getting documents from the Web requires an advance crawling techniques that is a research topic by itself. In this project, we are using other crawlers to collect the data from Web. Therefore, it is not relevant to measure the recall of CONIA. Instead, the aim of these experiments will be to measure the precision of the system.

The most of the experiments on IR systems, including the ones using TREC, base their measure of success to the first 10 returned documents for the query [11,52]. The main reason behind that is a user usually interested in checking the first few documents in the returned results to find an answer to his request.

In the following sections, we are going to evaluate queries for "Health" and "Terrorism" domains. The results of each query are evaluated by set of people. The following steps are used in the set up of the experiments.

- First, for a given description, a small survey is done for the determination of the query (Pre-survey). In this process, we have provided 1, 2 phrase description of what the user interested in finding in the documents and asked 3–5 people to write the relevant query for the given description. The result of this is assessed in two phases. If the majority of the people listed the query, that query is selected; otherwise, the most common words stated in the participants queries are selected and combined together.
- Once the query is known for the given description, it is passed to the search engine (Yahoo/Google) to retrieve the corresponding documents. In addition to the original query, we send several queries to the search engine to collect the results for the expanded set of terms of a query. For example, for the original query " eye disorder," the successive queries of " cataract " and "glaucoma" that are subclasses of " eye disorder" are also sent to the search engine.
- Mainly, Yahoo API is used to retrieve the documents for queries. Yahoo API has limit of 1,000 returned documents for each query. Even in some cases it shows that millions of documents are found, the upper limit of presented documents is 1,000. Although we haven't able to get permission to use Google API, for the experimental purposes we crawled the returned results of Google for a given query and used it in some of our experiments. Considering the successive queries of a given query with the expanded terms, CONIA corpus has an upper limit of couple of thousands. However, in order to reduce the response time of the system, we used 100 documents as an upper bound and CONIA's ranked results are returned from the list of 100 documents.

- The success (precision) of CONIA is measured by passing its top ranked results for user evaluation. Although that shows its individual success, it is also important to compare CONIA's success with the search engine used in forming the corpus. Therefore, we picked the first 7 returned results from search engine and CONIA, to pass them for the user evaluation. The reason of picking the first 7 results instead of 10 is to reduce the evaluation time required by the user about 30 %.
- For each query, the first 7 results of CONIA and the search engine are combined together in a single survey to pass to the users for evaluation. Since the results returned by the search engine changes in time by the new documents added to the Web, we saved the results of queries in a specific time with the returned links to the documents and provide these documents to all the users.

4.2 Metadata evaluation of surveys

Each survey passed to the users is designed to measure the success of a single query in CONIA and search engine. Each survey captures the following information provided by the user.

- User's profession and highest Degree.
- Given the problem, relevance of the selected query.
- For each document, the degree of relevance for the document (Not Relevant, Somehow Relevant, Probably Relevant and Relevant).
- Confidence level of user in his answers.

The queries used in the surveys are chosen to compare different aspects between CONIA and search engine. Overall, queries are organized to test the success of CONIA in following cases.

- CONIA in "Health" and "Terrorism" Domains.
- CONIA news-web (general search) articles.
- CONIA–Google comparison.
- CONIA–Yahoo comparison.
- CONIA–Google–Yahoo comparison for the same queries.

4.2.1 Survey statistics in health domain

This set of surveys represents the information request of a user in the health domain. The problem descriptions (dependently queries) of this domain are selected to cover the wide variety of topics as mentioned in NCI ontology. Similarly, survey participants are selected from medical school graduate students. In total, 15 queries are tested in 11 surveys, and one of the two interest ontologies is used to enhance each query. In 9 of the surveys, CONIA is compared with the one other search engine, and in the other 3 surveys, documents are combined to test the success of three systems Google, Yahoo and CONIA (Table 1).

Six queries are tested in two different surveys for the comparison purposes; 3 of these queries are used for comparing the documents retrieved from "news search" and "web search," and the other 3 are used to compare Yahoo/Google-news results against CONIA.

Although the same query tested in different cases, the returned results of these queries had very few or no overlaps. Thus, we are considering each case as an independent case.

In average, 8 people are participated in each survey and evaluated about 12 documents. Although the time spent for most of the users was about 5 min, the average time spent

Table 1 Health Domain queries and IR systems they are tested

	Description	Query	Comparison
1	Find the documents that provide you information about the health effects of air pollution on asthmatic people who are exercising	Air pollution exercise asthma	CONIA/YAHOO-web
2	Find the documents that provide you information about the health effects of air pollution on asthmatic people who are exercising	Air pollution exercise asthma	CONIA/YAHOO-news
3	Find the documents that provide you information about the children with brain tumor that had surgery	Brain tumor surgery children	CONIA/YAHOO-news
4	Find the documents that provide you information about the children with eye disorder that get blind	Eye disorder blind children	CONIA/YAHOO-news
5	Find the documents that provide you information about the children with eye disorder that get blind	Eye disorder blind children	CONIA/GOOGLE-news
6	Find the documents about the people who have had an heart attack from smoke and died	Heart attack smoke died	CONIA/YAHOO-news
7	Find the documents about the people who have had an heart attack from smoke and died	Heart attack smoke died	CONIA/GOOGLE-news
8	Find the documents that provide you information about the children who died from leukemia	Leukemia child died	CONIA/YAHOO -news
9	Find the documents about the children who had a seizure and become unconscious	Children seizure unconscious	CONIA/GOOGLE-news
10	Find the documents that talk about the unconsciousness of the children during seizure	Children seizure unconscious	CONIA/YAHOO-web
11	Find the documents that talk about the children that had kidney transplant	Kidney transplant children	CONIA/YAHOO-news
12	Find the documents that talk about the children that had kidney transplant	Kidney transplant children	CONIA/GOOGLE-news
13	Find the documents that talk about the death of infants with viral infections	Viral infection infant death	CONIA/YAHOO-web
14	Find the documents that talk about the death of infants with viral infections	Viral infection infant death	CONIA/GOOGLE-news
15	Find the children that had an immune system corruption	Immune system corruption children	CONIA/YAHOO-news

for each survey is measured as 12 that give less than a minute to analyze each document. The standard deviation of the participants in the "Health" domain surveys is calculated as 0.915.

Although less than a minute evaluation time for the document represents the real case for the "static informational search," it might not necessarily represent the ideal evaluation for

the "dynamic informational search," which misleads the evaluators in some cases. The reason is users with the static informational need usually look for the documents that are solely talking about the topic they are searching for. However, in the "dynamic informational search," the main goal of the user is not to get the definitional information. The information user's looking for mostly embedded in the small sections of the documents such as news. Therefore, considering the possibility of outliers, we compared different evaluation techniques in the following section.

4.2.2 Survey statistics in terrorism domain

This set of surveys represents the information request of a user in the "Terrorism" domain. The problem descriptions (dependently queries) of this domain are selected to cover the wide variety of topics as mentioned in "Terrorism" ontology. In two of the queries, the interest-ontology is used. Since it wasn't possible to find Intelligence Analysts to participate at surveys, the survey participants were selected from Computer Science graduate students. In total, 12 queries are tested by 11 surveys. Four of the queries are used twice by different search engines (Google and Yahoo) for the comparison purposes (Table 4).

On average, 9 subjects are participated in each survey and evaluated about 11 documents. Although the time spent by most of the users were about five minutes, the average time spent for each survey is measured as 9 min that is less than a minute for the analysis of each document. The standard deviation of the participants in the Terrorism domain surveys is calculated as 0.905 that is very close to the standard deviation of the user evaluations in Health domain.

4.3 Runtime complexity of CONIA

As much as the content of the returned results, the time spend on retrieving the requested information is an important factor for the users of the system. The success of CONIA is highly depending on the constructed ontology. The more information is available in the ontology, better the results of the system. Therefore, the experiment is made to estimate the overhead of the size of an ontology in the system. In the first experiment, the part of NCI ontology is used that consists of 15,258 concepts. In average, it took about 4 s to upload and do all the mappings on the ontology. Similarly, when the full version of NCI ontology is used (consists of 34,000 concepts) and the same queries are run, it is observed that it took in average 8 s to upload and do all the ontology-related calculations. Considering that this experiment is done on an average laptop with 2 GB of RAM and Intel Core i3 CPU 2.40 GHz, we observed that the size of ontology does not add too much overhead on the runtime of the program.

Similarly, an experiment is done to see how long it takes to retrieve and parse the documents. Since the information is not stored in the computer and the content of pages is retrieved in the current time, the time required to retrieve the content of the information is completely depends on the speed of the network. In addition, the time required to parse the retrieved document is related to the size of the document. The tests are done on 3.72 Mbps bandwidth. Although, it is not possible to give an exact measure for this testing, because of the above variants, the experiments showed that it takes an average 2 s to retrieve and process each document.

Overall, these experiments are done in the average conditions. We believe that by increasing the capabilities of the computer and the bandwidth of the internet connection, the amount of time needed to retrieve the required results can be minimized.

4.4 Evaluations of the queries

This section presents the evaluation results for the documents returned by each query. Among the evaluation methods used in measuring IR Systems performance, Normalized Discount Cumulative Gain (NDCG) [23] is used to measure the precision of CONIA. One of the main reasons of using NDCG is because P@n [44] and MAP evaluation methods only handle cases with binary judgments such as "relevant" and not relevant. NDCG is formulated to take into account the multiple level of relevance in addition to the ranking of the documents in calculating the relevance. Liu [29] used NDCG scores to calculate the success of the queries used in OHSUMED [20] subset. They transferred 3 levels of ratings "irrelevant," "partially relevant" and "definitely relevant" into the numeric numbers {0, 1, 2} to be able to use the values in the calculation of scores. Similarly, we used four-point scale {0, 1, 2, 3} to represent the ranking of the users' evaluation that is in the form of "not relevant," "somehow relevant," " probably relevant" and "relevant" and used the values in calculation of scores.

Because of the high standard deviation between the users' evaluations on the query documents, we used variations of evaluation techniques in extracting the user-assigned relevance of the documents. First, participants' overall judgment for each document is calculated and passed to NDCG for evaluation. The process of extracting the overall user-assigned relevance of the documents can be done in various ways as follows:

- **Mean-4PT**: The mean of the survey participants' evaluation at 4-pt scale {0, 1, 2, and 3} is taken for each document and passed to NDCG.
- **Mean-3PT:** The four-point evaluation of the survey participants is first transferred to 3-pt scale{0, 1, 2} by assigning "somehow relevant "and "probably relevant" judgments to value 1, which followed by passing the 3-pt Mean value to NDCG. The 3-pt scale is used to combine the evaluation results of the users that did not get clear distinction of whether the document belongs to "somehow relevant" or "probably relevant."
- **Mean-2PT:** The original evaluation (4pt) of survey participants is transferred to 2-pt {0, 1} scale. The "Not Relevant" and "Somehow Relevant" judgments are assigned to 0, while "Probably Relevant "and "Relevant" judgments are assigned to 1. Accordingly, the mean of participants' evaluation over 2-pt scale is passed to NDCG for evaluation. The 2-pt scale is used to compare the user feedback with the traditional binary judgments "not relevant" and "relevant."
- **Median-4PT:** The high standard deviation of the user evaluations usually an indication of the skewness in the data. For the data set with outliers, median is considered as a more robust measure [13]. Therefore, we used median value of the participants' evaluations (in 4pt scale) to pass NDCG for evaluations.

Table 2 shows the NDCG evaluation results, for 15 queries obtained from the 11 surveys of "Health" domain. The success of CONIA with Median-4PT user evaluation in the given set of queries changed between 30–95 % with the mean average of 62 %.

In order to better test the success of the CONIA, in Table 3, we compared its success with the other search engine/s that is used to retrieve the documents for its corpus. The comparison of the systems is done by using NDCG evaluation for the measures. In addition, the last 3 columns of the table show the number of documents considered as "relevant" and "irrelevant" in binary evaluation. The binary evaluation mentioned in the table is calculated using mean of the user's evaluation in which the documents with overall score of less than 0.5 are assigned to "irrelevant."

The overall success of CONIA in the "Health" domain for the first 7 documents of the given queries compared to the other IR systems is calculated between 12 and 17 % depending

Table 2 Ranking performance of CONIA in health domain

Query	CONIA'S success (NDCG @7)
Air pollution exercise asthma	0.750
Air pollution exercise asthma	0.602
Brain Tumor Surgery Children	0.565
Eye disorder blind children	0.519
Eye disorder blind children	0.575
Heart attack smoke died	0.317
Heart attack smoke died	0.626
Leukemia child died	0.406
Children seizure unconscious	0.663
Children seizure unconscious	0.955
Kidney transplant children	0.757
Kidney transplant children	0.654
Viral infection infant death	0.754
Viral infection infant death	0.520
Immune system corruption children	0.552
Average	0.618

on the calculation of the user-assigned relevance. Although success of CONIA is calculated for different user-assigned relevance measures (Table 3), since the original surveys are conducted using 4 point scale, we used 4-pt scale as a main comparison scale in the rest of the experiments. The difference between the 4-pt Mean and Median in Table 3 is calculated as 0.3(11.5, 11.8 %). Although this is a very minor difference, since Median is accepted as a more robust measure than the Mean in a data set with skewed information [13], in the rest of the discussion, the main success of CONIA will be measured in terms of Median.

As a result of 15 query evaluation, CONIA failed in 3 queries and get a negligible success on the other compare to the used search engine results. Three of the 4 queries showed around 2 % difference than the compared search engine, which is negligibly small. However, in the query "Brain Tumor Surgery Children," the used search engine performed 22 % better than CONIA. The reasons behind CONIA's failing for this query is investigated and found that the document ranked first in CONIA evaluated less relevant than the one found by Yahoo, which caused the dramatic difference in the results. NDCG evaluation takes into account both the evaluation value and its ranking in the document set. If the document with high rank is misjudged, the evaluated NDCG value becomes much lower from the case if the misjudgment is lower ranked.

The normalization of Discount Cumulative Gain (DCG) is calculated by dividing DCG with the DCG value of the 7 most relevant documents chosen by the users. This normalization is mainly based on the assumption that survey participants evaluate all the possible documents in the given set and chose the 7 best documents representing the given query.

The total number of "relevant" documents found, among the first 7 most relevant set of documents, by the users of the 15 queries of "Health" domain is calculated as 55. Therefore, the success of CONIA and the search engine (Yahoo/Google) finding the relevant documents in their first 7 returned documents (P@7) in binary scale calculated as 34(62 %) and 26(47 %).

Similarly, Table 4 shows the NDCG evaluation results, for 12 queries obtained from the 10 surveys of "Terrorism" domain for CONIA when Median-4PT user evaluation is used to

Table 3 Success of CONIA compared to (Google/Yahoo) by using NDCG evaluation measure

Query	Comparison	4PT-Mean	3PT-Mean	2PT-Mean	4PT-Median	User_R	Con_R	Yahoo_R/Google_R
Air pollution exercise asthma	CONIA/YAHOO-web	0.09	0.01	0.05	0.08	7	5	5
Air pollution exercise asthma	CONIA/YAHOO-news	0.24	0.12	0.16	0.23	3	1	0
Brain tumor surgery children	CONIA/YAHOO-news	−0.02	−0.05	0.09	−0.22	5	4	3
Eye disorder blind children	CONIA/YAHOO-news	0.03	0.01	0.11	−0.02	3	1	1
Eye disorder blind children	CONIA/GOOGLE-news	0.08	0.06	0.07	−0.02	3	1	2
Heart attack smoke died	CONIA/YAHOO-news	0.42	0.47	0.40	0.17	1	0	0
Heart attack smoke died	CONIA/GOOGLE-news	−0.01	−0.02	0.04	0.01	1	1	1
Leukemia child died	CONIA/YAHOO-news	0.23	0.12	0.27	0.28	3	2	1
Children seizure unconscious	CONIA/GOOGLE-news	0.15	0.17	0.09	0.14	6	3	2
Children seizure unconscious	CONIA/YAHOO-web	−0.29	0.21	0.30	0.46	5	4	2
Kidney transplant children	CONIA/YAHOO-news	0.10	0.03	0.13	0.22	3	2	1
Kidney transplant children	CONIA/GOOGLE-news	0.05	−0.02	0.06	0.18	3	3	3
Viral infection infant death	CONIA/YAHOO-web	0.06	0.05	0.07	0.07	4	2	1
Viral infection infant death	CONIA/GOOGLE-news	0.07	0.05	0.07	0.10	4	2	2
Immune system corruption children	CONIA/YAHOO-news	0.04	−0.01	0.05	0.08	4	3	2
Average		0.12	0.08	0.17	0.12			

Table 4 Ranking performance of CONIA in terrorism domain

Query	CONIA'S success (NDCG @7)
Cyber attack United States military	0.613
Cyber attack United States military	0.762
Iraq girl kidnapped terrorist	0.771
PKK bombing politics	0.881
Plane Hijack United States	0.757
Plane Hijack United States	0.616
Terrorist attack United States economy	0.807
Explosion government building	0.757
Explosion government building	0.800
Child abduction California	0.763
Firearm shooting political events	0.880
Firearm shooting political events	0.857
Average	0.772

assign the relevance measures of the documents. The success of CONIA in the given set of queries changed between 61 and 88 % with the mean average of 77 %.

As in the "Health" domain, success of CONIA in "Terrorism" domain is compared to the other search engine/s that is used to retrieve the documents for its corpus (Table 5). The overall success of CONIA compared to other search engine/s in the Terrorism domain has been found as 27 % in Median calculation.

The total number of "relevant" documents found, among the first seven most relevant set of documents, by the users for the seven queries in Terrorism domain is calculated as 64. Therefore, the success of CONIA and the search engine (Yahoo/Google) on finding the relevant documents in their first seven returned documents (P@7) in binary scale calculated as 48(75 %) and 33(52 %).

As mentioned earlier, the surveys are designed to test several cases. One of the cases is the search type (news search and Web search) used to collect the documents for the corpus of CONIA. When the results of the both "Health" and "Terrorism" domains combined together, the compared success of CONIA with the corpus from "web search" is measured 25 % for the 9 queries. Similarly, the compared success of CONIA for the overall "news search" is measured as 15 % for the 18 queries. The difference in the success rate we believe comes from the case that 2/3 of the queries in the "web search" was from "Terrorism" domain when compared to the "news search" where only 6/18 queries were from "Terrorism" domain. We also had two queries to be tested in both types, but the results were inconclusive and we believe more experiments need to done for the comparison of this case to come up with a conclusion.

In addition, CONIA is also compared to Yahoo and Google for its overall success. Out of 27 queries in total, Google is used to retrieve the documents for 11 queries. In these 11 queries, the success of CONIA compare to Google is measured as 13 % by Median. For the 16 queries that Yahoo is used to crawl the documents, CONIA's success is measured as 19 % (Table 6).

In overall, 6 queries (query set 1) are tested on both Google and Yahoo with the overlap of the common documents between Google and Yahoo as 12 %. The success of CONIA in these 6 queries compared to Google and Yahoo is measured as 11 and 21 % (Table 6).

4.5 Summary of evaluation results

In this paper, 27 queries are tested through 21 surveys. The evaluation results showed that CONIA's average success is measured as 62 and 77 % in "Health" and "Terrorism" domains. When CONIA's success is compared to the search engine (Google/Yahoo) used for crawling the data for CONIA's corpus, in overall CONIA performed 12 and 27 % better than the other search engine/s in "Health" and "Terrorism" domains. The reason behind the different success rates between the domains mainly came from the fact that "Health" domain had less relevant set of document compare to the "Terrorism" domain. Participants of the "Health" domain survey rated most of the documents negatively (i.e., most of the values tend to be between 0 and 1). The DCG evaluation method calculates the relevance of the documents based on their order and the evaluation value in which grows exponentially by the relevance value given to the document. Since the difference between the document evaluations of the users lied in a small interval dependently, the difference between the evaluation results also stayed small.

Our observation while working with the "news search" showed that news search in both Google and Yahoo is sensitive to time and does not include the documents that are older than a month. In some cases, this restriction limited the number of possible documents retrieved

Table 5 Success of CONIA in terrorism domain is compared to (Google/Yahoo) by using NDCG evaluation mean

Query	Comparison	4PT-Mean	3PT-Mean	2PT-Mean	4PT- Median	User_ R	Con_ R	Yahoo_R/ Google_R
Cyber attack United States military	CONIA/GOOGLE-news	0.01	0.04	−0.02	0.10	7	4	5
Cyber attack United States military	CONIA/YAHOO-news	0.30	0.29	0.13	0.30	7	5	4
Iraq girl kidnapped terrorist	CONIA/YAHOO-web	0.44	0.38	0.40	0.41	5	4	2
PKK bombing politics	CONIA/GOOGLE-web	0.22	0.08	0.33	0.22	7	6	4
Plane Hijack United States	CONIA/GOOGLE-news	0.23	0.22	0.23	0.33	3	3	1
Plane Hijack United States	CONIA/YAHOO-news	0.31	0.39	0.43	0.39	3	3	1
Terrorist attack United States economy	CONIA/GOOGLE-web	0.35	0.34	0.27	0.15	7	6	4
Explosion government building	CONIA/YAHOO-news	0.21	0.06	0.14	0.21	3	2	1
Explosion government building	CONIA/GOOGLE-news	0.12	−0.02	0.11	0.17	3	2	3
Child abduction California	CONIA/YAHOO-web	0.18	0.18	0.05	0.13	6	4	2
Firearm shooting political events	CONIA/YAHOO-web	0.21	0.22	0.08	0.36	6	4	3
Firearm shooting political events	CONIA/GOOGLE-web	0.25	0.22	0.10	0.41	7	5	3
Average		0.24	0.20	0.19	0.27			

Table 6 Success of CONIA compared to Google and Yahoo on the same data set shared by all

	4PT-Mean	3PT-Mean	2PT-Mean	4PT-Median
Yahoo–CONIA Query set 1	0.23	0.21	0.22	0.21
Google–CONIA Query set 1	0.07	0.04	0.08	0.11
Yahoo–CONIA overall	0.20	0.16	0.17	0.19
Google–CONIA overall	0.14	0.10	0.12	0.13

Success of CONIA is compared to Yahoo and Google through all the surveys

for the corpus of CONIA and restricted us to certain queries since not all the queries provided result. Collecting data from Web search did not have the scarcity problem as of news search, but the documents retrieved by "web search" tend to be more "static information" type than the "dynamic information "that contradicts the main interest of the users of CONIA. In total, Google showed about 6 % better success rate than Yahoo, and CONIA outperforms Google's success by 13 %.

The overall results of 21 surveys proved that using semantic information extracted from lexical and domain ontologies in the context of domain users for mapping into the documents improves the performance of the Information Retrieval Systems noticeably.

References

1. Auger A, Morin M-A (2005) TerroGate: a new information extraction technology designed for the terrorism domain. Défense Sécurité Innovation, Quebec City, Québec
2. Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. Addison Wesley, Reading. ISBN 0-201-39829-X
3. Baldauf M, Dustdar S, Rosenberg F (2004) A survey on context-awere systems. Int J Ad Hoc Ubiquit Comput 2(4): 263–277
4. Baralis E, Cagliero L, Cerquitelli T, Garza P, Marchetti M (2011) CAS-MINE: providing personalized services in context-aware applications by means of generalized rules. Knowl Inf Syst 28(2):283–310
5. Bellahsene Z, Bonifati A, Duchateau F, Velegrakis Y (2011) On evaluating schema matching and mapping, 1st edn, XII, p 314
6. Broder A (2002) A taxonomy of Web search. SIGIR Forum. #36(2):3–10
7. Chen C-L, Tseng FSC, Liang T (2011) An integration of fuzzy association rules and WordNet for focumenting clustering. Knowl Inf Syst 28(3):687–708
8. Clarke CLA, Cormack GV, Tudhope EA (2000) Relevance ranking for one to three term queries. Inf Process Manag Kowalski 36(2):291–311
9. Falconer SM, Noy NF (2011) Interactive techniques to support ontology matching. Data-centric systems and applications. Springer, Berlin. doi:10.1007/978-3-642-16518-42
10. Fodeh S, Punch B, Tan P-N (2011) On ontology-driven document clustering using core semantic features. Knowl Inf Syst 28(2):395–421
11. Gao J, Walker S, Robertson S, Cao G, He H, Zhang M, Nie J-Y (2001) TREC-10 web track experiments at MSRA. In: NIST special publication : the tenth text retrieval conference (TREC). National Institute of Standards and Technology, Gaithersburg
12. Gerald S, Michael JM (2008) Introduction to modern information retrieval. McGraw-Hill, New York
13. Greisdorf H, Spink A (2001) Median measure: an approach to IR systems evaluation. Inf Process Manag Elisever 37:843–857
14. Griesbaum J (2004) Evaluation of three German Search engines: Altavista.de, Google.de and Lycos.de. Inf Res 9(4): 9–4.

15. Grunewald L, McNutt G, Mercier A (2003) Using an ontology to improve search in a terrorism database system. In: Proceedings of the 14th international workshop on database and expert system applications, DEXA
16. Guarino N (1998) Formal ontology in information systems. IOS Press, Amsterdam
17. Gulla JA, Auran PG, Risvik KM (2002) Linguistics in large-scale web search. In: Proceedings of the 7th international conference on applications of natural language to information systems NLDB, Stockholm
18. Gupta DK (2005) Exploring roots of terrorism. In: Bjørgo T (ed) Root causes of terrorism. Routledge, London
19. Hawking D, Thistlewaite P (1996) Relevance weighting using distance between term occurrences, Unpublished manuscript, joint computer science technical report series, The Australian National University
20. Hersh W, Buckley C, Leone T, Hickman D (1994) Ohsumed: an interactive retrieval evaluation and new large text collection for research. In: Proceedings of SIGIR-94, 17th ACM international conference on research and development in information retrieval, Dublin
21. Henzinger MR, Motwani R, Silverstein C (2002) Challenges in web search engines. SIGIR
22. Houghton B (2002) Understanding the terrorism database. National Memorial Institute for Prevention of Terrorism Quarterly Bulletin
23. Jarvelin K, Kekalainen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20(4):422–446
24. Jiang J, Conrath, D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings on international conference on research in computational linguistics, Taiwan
25. Kowalski G (1997) Information retrieval systems, theory and implementation. Kluwer Academic Publishers, Boston
26. Krebs VE (2001) Mapping networks of terrorist cells. Connections 24(3):43–52
27. Liu S, Liu F, Yu CT, Meng W (2004) An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval
28. Liu B (2006) Personal evaluations of search engines: Google, Yahoo! and MSN. Department of Computer Science University of Illinois at Chicago
29. Liu T, Xu J, Qin T, Xiong W, Li H (2007) LETOR:benchmark dataset for research on learning to rank for information retrieval. In: SIGIR '07 workshop on learning to rank for information retrieval
30. Long H, Lv B, Zhao T, Liu Y (2007) Evaluate and compare Chinese internet search engines based on users' experience. In: Proceedings of IEEE wireless communications, networking and mobile computing conference, WiCom
31. Lopez V, Victoria U, Marta S, Enrico M (2009) Cross ontology query answering on the semantic web: an initial evaluation. In: Proceedings of the 5th international conference on knowledge capture
32. Mannes A, Golbeck J (2007) Ontology building: a terrorism specialist's perspective. In: Proceedings of the IEEE Aerospace conference
33. Michelizzi J (2005) Semantic relatedness applied to all words sense disambiguation. Master's thesis, University of Minnesota, Duluth
34. Miniwatts Marketing Group, InternetStat (2009). www.internetworldstats.com/stats.htm
35. Moldovan DI, Mihalcea R (2000) Using WordNet and lexical operators to improve internet searches. IEEE Internet Comput 4(1):34–43
36. Morgan K (1992) MUC-4 proceedings. In: Proceedings of the fourth message understanding conference (MUC-4), San Mateo
37. Patwardhan S, Banerjee S, Pedersen T (2003) Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the fourth international conference on intelligent text processing and computational linguistics
38. Pedersen T, Patwardhan S, Michelizzi J (2004) WordNet::similarity—measuring the relatedness of concepts. In: Proceedings of the nineteenth national conference on artificial intelligence (AAAI)
39. Pingdom R (2009) Internet 2009 in numbers. http://royal.pingdom.com/2010/01/22/internet-2009-in-numbers/
40. RAND Corporation (2003) Purpose and description of information found in the incident databases . http://www.tkb.org/RandSummary.jsp
41. Rasolofo Y, Savoy J (2003) Term proximity scoring for keyword-based retrieval systems. In Proceedings of the 25th European conference on IR research
42. Roberto N, Litkowski KC, Orin H (2007) Coarse-grained english all words task. In Proceedings of the 4th international workshop on semantic evaluations semEval, Prague
43. Rose DE, Levinson D (2004) Understanding user goals in web search. In WWW '04: Proceedings of the 13th international conference on World Wide Web, ACM, New York

44. Salton G, McGill M (1983) An introduction to modem information retrieval. McGraw-Hill, NewYork, NY
45. Singhal A, Buckley C, Mitra M (1996) Pivoted document length normalization. In: ACM-SIGIR Conference on research and development in information retrieval
46. Smeaton AF, Berrut C (1996) Thresholding postings lists, query expansion by word-word distances and POS tagging of Spanish text. In: Proceedings of the fourth text retrieval conference
47. Smith BL, Damphousse KR (2002) The American terrorism study: indictment database
48. Spink A, Wolfram D, Jansen BJ, Saracevic T (2001) Searching the web: the public and their queries. J Am Soc Inf Sci 53(2):226–234
49. Tao T, ChengXiang Z (2007) An exploration of proximity measures in information retrieval, SIGIR
50. US Environmental Protection Agency (2009). http://www.epa.gov/air/urbanair/
51. Ussery B (2008) Average number of words per query have increased! http://www.beussery.com/blog/index.php/2008/02/google-average-number-of-words-per-query-have-increased/
52. Webber W, Moffat A, Zobel J (1983) Score standardization for intercollection comparison of retrieval systems (SIGIR)
53. Weinberger H (2011) Search in context. Lecture notes in business information processing, enterprise information systems
54. Zhang Z (2005) Ontology query languages for the semantic web: a performance evaluation. Master's thesis, University of Georgia

Author Biographies

Vesile Evrim is currently an Assistant Professor in the Department of Information Systems Engineering at the Cyprus International University. She received her Ph.D. and M.S. degrees in Computer Science from University of Southern California in 2009 and 2003 and M.S. and B.S. degrees in Applied Mathematics and Computer Science from Eastern Mediterranean University in 1999 and 2001. Her research interests include customized information retrieval, ontologies, text-based data mining, information trust, recommender systems and social networks.

Dennis McLeod is currently Professor of Computer Science at the University of Southern California, and Director of the Semantic Information Research Laboratory. He received his Ph.D., M.S., and B.S. degrees in Computer Electrical Engineering from MIT. Dr. McLeod has published widely in the areas of data and knowledge base systems, federated databases, database models and design, ontologies, knowledge discovery, scientific data management, information trust and privacy and multimedia information management. His current research focuses on structured domain ontologies; semantic web; database semantic heterogeneity resolution and inter-database correlation; personalized information management and customization; information management environments for Earth, marine, and climate science; the architecture of data centers providing massive storage via virtualization and data clouds; social networking information management and information trust; and service-based information access and delivery frameworks.