A MULTILEVEL UNL CONCEPT BASED SEARCHING AND RANKING

E. Umamaheswari, T. V. Geetha, Ranjani Parthasarathi and Madhan Karky
College of Engineering, Anna University, Chennai, India

Keywords: CoReS, UNL, Concept relation search and ranking, Query Expansion, Query Translation.

Abstract: The recent advances in search engines have resulted in a huge explosion of available web documents. Understanding the content on the web and providing meaningful search results to the user have become essential for any search engine. This paper proposes CoReS, a multilevel Concept based Searching and Ranking Algorithm which retrieves and ranks the documents based on the concepts and relationships between the concepts. The search and rank methodology is based on Universal Networking Language (UNL) representation of the documents. The UNL Index based query expansion technique is used to provide more meaningful results to the user. The algorithm has been evaluated on a corpus of tourism documents, and its performance compared with keyword based search. The mean average precision of the concept based search is found to be 0.75 while the keyword based search has a MAP score of 0.45.

1 INTRODUCTION

The content on the web is growing rapidly every fraction of a second. Search engines such as Google, Yahoo and MSN have become the most heavily-used online services, with millions of searches performed every day. All the above search engines basically use keyword based search strategy. The ranking algorithms such as PageRank algorithm(Brin and Page, 1998) and HITS Algorithm(Brin and Page, 1998) score the documents according to the incoming and outgoing links of the documents. However due to the huge number of documents available on the web, the number of results produced by keyword based search engines is too many. The ultimate challenge for search engines is to provide effective systems that retrieve the most relevant information from the web that exactly caters to the users information need.

Concept based search attempts to improve search effectiveness by incorporating conceptual information that convey meaning rather than using the presence or absence of keywords as the basis for the retrieval process. Concept based search can be classified as those that use a background knowledge source to provide conceptual information and those that use semantically analyzed components of the document. Concept based search can also be classified based on how semantics is used to represent the documents. Documents can be represented by considering concepts associated with the frequently occurring keywords or by converting important components of the document into a semantic structure. In addition, concept based search can also be classified based on where the semantics is introduced in the components of the search engine. Semantics can be introduced in query expansion, building the index, searching and also in ranking the search results.

This paper deals with a concept based search which uses a semantic representation of documents, and incorporates semantics in all the components of the search engine. Universal Networking Language(M and H, 1998) (UNL), an inter-lingual language independent semantic representation is used for document representation. UNL based concepts and relations are used to build the index structure; and this UNL based index is used for query expansion, search and ranking.

The focus of this paper is on the algorithm, CoReS, used for UNL based Concept relation search and ranking. This algorithm aims at improving the search and ranking by performing matching at three levels, namely

1. Partial or Complete match between the index and expanded query;
2. Concept Association level which distinguishes between actual query terms,query concepts and expanded concept; and

Umamaheswari E., V. Geetha T., Parthasarathi R. and Karky M..
A MULTILEVEL UNL CONCEPT BASED SEARCHING AND RANKING.
DOI: 10.5220/0003334402820289
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 282-289
ISBN: 978-989-8425-51-5
Copyright © 2011 SCITEPRESS (Science and Technology Publications, Lda.)

3. Document based features such as frequency of occurrence and position of terms and concepts in the document.

This paper is organized as follows. Section 2 of this paper gives an overview of related work in concept and semantic based search. Section 3 describes overall architecture of the concept based search system and describes the UNL structure. It also outlines the various features available in the index that are used for matching and ranking; as well as the features included in the UNL based query expansion and translation. Section 4 describes CoReS, a Multilevel UNL Concept relation based Searching and ranking. Section 5 discusses the results and the evaluation of the UNL Searching and Ranking, and Section 6 presents the conclusion and future work.

2 RELATED WORK

This section explores literature related to concept based searching and ranking. Only a few meaning based search engines have been developed.

2.1 Semantic Search Engines

While some meaning based search engines use sentence level semantics, others use ontology as the background knowledge source for providing semantics. Hakia(HAKIA, 2009) is a semantic search engine that uses knowledge of Ontology and Fuzzy logic for semantic ranking. In order to retrieve conceptual results, it uses QDEX (Query Detection and Extraction) Indexing Architecture which enables semantic analysis of web pages and provides meaning based search results. In Hakia besides the keywords, phrases are used for meaning based searches.

The limitation of Hakia is that it accepts queries as questions in a specific format. Also, the QDEX algorithm extracts all possible queries that can be asked on the content of web pages of various lengths and forms. This is an offline process before any user query is entered. The major difficulty in QDEX system is the reduction of the huge number of generated query sequences into a few dozens that make sense. Hakia allows only these predefined query sequences generated from the content to be used as queries.

On the other hand SenseBot(Sensebot, 2009) is a semantic search engine that runs over search engines like google and yahoo to generate multi document summary based on text mining and limited semantics.

Though all the above search engines provide meaning based results, some search engines require sophisticated query analysis techniques to provide meaningful search results. Other search engines consider concepts rather than relations between concepts as the basis of match. However, in the search engine described in this paper, the context of the query is retrieved by traversing the already created UNL based indexer. The frequently occurring UNL relations obtained from the UNL index, in effect provide information about the possible connections between concepts in the specific domain under consideration. These connections provide the context of the query concept, and the query expansion based on this context yields meaningful search results.

2.2 Ontology based Semantic Search Engines

Concept based search can also be based on the use of knowledge structures. One such search engine is Engineering or Environmental Knowledge Ontology-based Semantic Search(EKOSS)(Kraines et al., 2006). It is an ontology based semantic search engine which uses a fully functional ontology for representing the knowledge base. It provides a collaborative knowledge sharing environment and helps knowledge experts to share their knowledge such as research papers, database, computer simulated model and even curriculum vitae. The EKOSS system is used to construct computer-interpretable semantically rich statements of the knowledge resource. When a user request is posted, this system converts the user request into a computer readable knowledge description based on description logic and associated rules.

Ontology-based information retrieval(Gao et al., 2005) intended for e-Government has been developed for securing the legal documents of the government. The disadvantage of using ontology based search engines is that they are susceptible to changes in the information resources. This will affect the conceptualization of the domain representation. More over, the effort required to build an ontology is huge. This task is domain dependent and the use of common vocabulary ontology for different domains remains a challenging task.

2.3 UNL based Search Engines

A meaning based multilingual search engine that uses UNL (Universal Networking Language) is AgroExplorer(Surve et al., 2004). This search engine is similar to the search engine described in this work, since AgroExplorer also uses Universal Networking Language (UNL) expressions for representing sentences as graphs that capture the meaning of the sentences. The System has been developed for agriculture domain and also provides multilingual feature. It uses a simple search and rank process based on the degree of match of the query UNL and the frequency of occurrence of the Concepts with other concepts in the UNL expression.

The algorithm for searching and ranking described in this paper, is a part of UNL search system that differs from the existing AgroExplorer(Surve et al., 2004) in that it incorporates semantics in every component of the search engine. Also, it uses a sophisticated three-level search and rank process and a context-based query expansion to enhance the results obtained for a search.

3 BACKGROUND

This section gives a brief introduction to UNL (Universal Networking Language) and describes the UNL index structure and the UNL based Query expansion for conceptual searching and ranking.

3.1 The Universal Networking Language

The Universal Networking Language has been introduced as a digital meta-language for describing, summarizing, refining, storing and disseminating information in a machine-independent and human-language-neutral form (UNDL, 2009).

Words are expressed as concepts called as Universal Words or UWs. The UWs can be linked together with a relation. Relations specify the role of words in the sentences. There are a standard set of 46 UNL relations. The subjective meaning intended by the speaker can be expressed through attributes. Normally, natural language sentences are converted to UNL graphs or expressions using linguistic analysis. In the UNL expression, nodes represent concepts, and arcs represent relations between concepts.

The Knowledge Base (UNLKB) is provided to define the semantics of UWs. The UNLKB defines hierarchical relations and inference based relations between concepts.

3.2 UNL Index for Conceptual Search

In the UNL based search system discussed here, UNL graphs that represent fragments of sentences in a document are used to build the conceptual index. The UNL enconverter of the system uses a rule based approach to convert the sentence constituents to UNL graphs where concepts are represented as nodes and relations as edges. The use of this approach allows terms to be represented as concepts, extracts out a standard set of semantic relations between concepts in a sentence, and at the same time, associates a hierarchy for the concepts linked through the UNL semantic relations. This essentially means that semantically analyzed information from the sentences of the documents is used for building the index. In addition, the constraints associated with the concepts available in the UNL KB also incorporate information from a background knowledge resource into the index structure.For example the UW word *Chennai* of the tamil sentence will be translated into $Chennai(icl > place)$.Here *Chennai* denotes the head word and the $icl > place$ denotes the contraints associated with the concept. Figure 1 shows the UNL enconversion process.

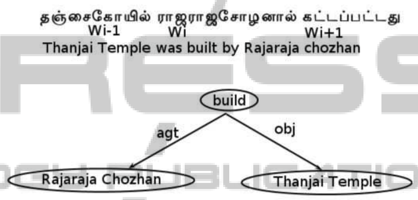

Figure 1: UNL Enconversion of a Tamil Sentence.

In $Figure1$ the concept $build(icl > action)$ is connected to $Rajarajachozhan(icl > person)$ and also with $ThanjaiTemple(iof > temple)$ using agt and obj respectively.

The set of UNL graphs obtained from the enconversion component of the search system are represented as a multi-list structure. This multi-list structure is converted into three separate indices CRC (Concept-Relation-Concept), CR (Concept-Relation) and C (Concept) indices in order to aid searching and ranking. In addition to building the UNL graph represented as multi-list structure, the UNL enconverter also provides additional information to aid the retrieval process.

The CRC Indices for the UNL tamil sentences are

1. $build(icl > action)$-agt-$Rajarajachozhan(icl > person)$
2. $build(icl > action) - obj - ThanjaiTemple(iof > temple)$

The CR Indices for the UNL tamil sentences are

1. $build(icl > action)$-agt
2. $build(icl > action)$-obj

The C Indices for the UNL tamil sentences are

1. $build(icl > action)$
2. $Rajarajachozhan(icl > person)$

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

3. $ThanjaiTemple(iof > temple)$

Sentence based information includes sentence identifier, Part Of Speech tags, Entity tags, Multiword tags, the actual terms or words associated with the UNL concepts and a bit pattern vector that indicates sentence-wise position of the concepts in the document. Document based information includes document identifier, term frequency, concept frequency, and the position of the concepts in the document. These features are used in weight determination during searching and ranking of documents. Features such as frequency of concepts present in the document in addition to term frequency, allow ranking to be both term and concept based which becomes important when term frequency is not significant. The bit pattern vector indicating distance between concepts helps to identify relations that are not necessarily proximity dependent.The UNL index with all the above sentence level and document level informations are stored in the Binary Search Tree(BST).

3.3 Context based Query Expansion

An important contribution of this paper is the use of semantics in the query expansion component of the search engine. In this work, context of a query concept is defined as the association of this concept with other concepts in a CRC relation, across documents in the domain of interest. By analyzing the index, the concept associated with a query is matched with the CRCs of the index and the most common CRCs associated with the query concept are extracted. The expanded concepts obtained, are ranked based on frequency of CRC and on its being an entity. Query expansion is an on-line activity and the index analysis results in efficient query expansion. The most frequently occurring CRC in the index indicates the frequent association of concepts in the domain across documents and hence gives the domain context of the query concept. This expansion of the query concepts to CRC allows context dictated query sub graphs to be constructed for the query. The expanded query graph is now associated with actual query terms, query concepts and expanded concepts associated with the context of the query concept. This in turn means that differentiation between these is required during both searching and ranking.

The index based query expansion influences the searching and ranking of documents in many ways. The association of expanded concepts with the query, helps to build CRC query graphs that can be matched with the UNL index. Without this expansion, single word queries would have resulted in isolated concept (C) only match while with the expansion we are matching with a context dictated CRC. As already explained, the association of expanded concepts allows domain oriented, corpus based context of the query word to play a role in semantic matching and in addition helps to bring in documents which have concepts in the context of the query, which would have been missed by other search mechanisms.

4 CONCEPTUAL SEARCHING AND RANKING

The basic searching procedure is based on complete CRC Match or partial CR or C matches between query sub graphs and the corresponding index as in AgroExplorer(Surve et al., 2004). However, in this paper, the design of the ranking procedure depends on whether the match of the index is with the actual query terms, actual query concepts or expanded concepts. In addition, all the sentence and document based features associated with the conceptual indices also affect the ranking procedure.

The overall algorithm for searching and ranking actually performs three level ranking. The first level ranking is obtained based on whether there is complete match (CRC match), partial match of Concept Relation (CR) or match of only concepts (C Only). This level of ranking is provided by the Degree of Match Categorization tag Ta. The set of documents obtained in level 1 category is further prioritized using Concept Association Categorization Tag Tb. Concept Association categorization depends on whether the index match is between query terms, query concepts or expanded concepts. Once the documents have been ranked by Ta and Tb, the documents at the same $Ta.Tb$ level are ranked based on weights calculated based on the index based features associated with the concept.

A Tag represented as $Ta.Tb$ helps in determining the two level list of prioritized documents. Tag Ta computed in level 1 indicates degree of match while Tb computed in level 2 indicates the type of concept association. For determining the tags the following terminology is defined.

A given query with n terms may be represented as a set Q,Let $Q = \{q_1, ..., q_n\}$,where q_i represents a query term. Each element i of the power-set of Q is expanded and enconverted to a set EQ_i of UNL graphs g_{im} ,where m represents the expanded concepts from the UNL index and $m > 0$.Here the power set of Q represents that each query term is associated with not only a single expanded terms and it's concepts,it also represents more than one expaned terms and concepts.

That is,$EQ_i = \{g_{i1}, g_{i2}, ...g_{in}\}$,where each g_{ij} is a tuple of $\{Cx_{ij}, R_{ij}, Cy_{ij}\}$ representing a relation R_{ij} between the two associated concepts X_{ij} and Y_{ij}. The presence of all three elements of the tuple corresponds to a *CRC* graph, the presence of a *C* and *R* corresponds to a *CR* graph, and the presence of a *C* alone indicates a *C* graph.Now each g_{ij} is matched with the *CRC*,*CR* and *C* indices represented in the index graphs in the indices I_{CRC}, I_{CR} and I_C to obtain a set of documents $D_{ij} \cdot_{CRC}$, $D_{ij} \cdot_{CR}$, and $D_{ij} \cdot_{C}$. The matching set of documents D_{ij} for the expanded query graph g_{ij} is the union of these three sets. i.e.
$D_{ij} = D_{ij} \cdot_{CRC}$ U $D_{ij} \cdot_{CR}$ U $D_{ij} \cdot_{C}$

Now by using these sets, the degree of match is determined by the tag *Ta*.

4.1 Tag Determination for Degree of Match

The tag determination for the degree of match depends on the extent of match between the CRC representing the query sub graph and the conceptual index. It essentially differentiates between CRC, CR and C matches. Ta helps in differentiating between the different degrees of match. The UNL sub graph is a directional graph and hence partial match also considers whether the concept in CR (Concept Relation), matches with the source concept, C_{x_i}, or destination concept, C_{y_i}, of the UNL subgraph.

$$Ta = \left\{ \begin{array}{ll} 1 & if\ \ q_i\{C_{x_i}, R_i, C_{y_i}\} \in I_{CRC} \\ 2 & if\ \ q_i\{C_{x_i}, R_i\} \in I_{CR}\ \ and\ \ q_i\{C_{y_i}, R_i\} \in I_{CR} \\ 3 & if\ \ q_i\{C_{x_i}, R_i\} \in I_{CR}\ \ and\ \ q_i\{C_{y_i}\} \in I_C \\ 4 & if\ \ q_i\{C_{y_i}, R_i\} \in I_{CR}\ \ and\ \ q_i\{C_{x_i}\} \in I_C \\ 5 & if\ \ q_i\{C_{x_i}\} \in I_C\ \ and\ \ q_i\{C_{y_i}\} \in I_C \\ 6 & if\ \ q_i\{C_{x_i}\} \in I_C \\ 7 & if\ \ q_i\{C_{y_i}\} \in I_C \end{array} \right\}$$

As shown above, the tag Ta has seven values bringing out the degree of match. Let $D_{ij}.Ta$ be the set of matched documents associated with each *Ta*.

4.2 Tag Determination for Concept Association

The next level of tag determination is based on whether the *Ci* value in *CRC*,*CR* and *C* matches corresponds to the actual query term ,the concept of the query term or the concept obtained after query expansion. Accordingly the concept association is said to be of three types.

1. Query Term TWi association - This means that the concept *Ci* is query term itself
2. Concept Word CWi association - This means that the concept *Ci* matches the corresponding concept of the query,but the actual query term is different.
3. Expanded Word EWi association - This means that the concept *Ci* is associated with a concept that is not actually in the query but has been obtained as a result of query expansion.

Based on the above 3 values the eight different tags are obtained as given below

$$Tb = \left\{ \begin{array}{ll} 1 & if\ \ C_{x_i} = C_{y_i} = TW \\ 2 & if\ \ C_{x_i} = TWandC_{y_i} = CW \\ 3 & if\ \ C_{x_i} = CWandC_{y_i} = TW \\ 4 & if\ \ C_{x_i} = C_{y_i} = CWs \\ 5 & if\ \ C_{x_i} = TWandC_{y_i} = EW \\ 6 & if\ \ C_{x_i} = EWandC_{y_i} = TW \\ 7 & if\ \ C_{x_i} = CWandC_{y_i} = EW \\ 8 & if\ \ C_{x_i} = EWandC_{y_i} = CW \end{array} \right\}$$

It can be seen that the Tag Tb,differentiating between the three types explained above, also differentiates between whether the concept is the source node *Cx* or destination node *Cy* of the directed UNL subgraph. The eight values of Tb bring out these differences.With in each DTa the documents are ordered as per Tb.

Each of the set of $D_{ij}.Ta$ documents are now tagged with the Tb tag. In other words, the searched documents are prioritized and ranked according to $Ta.Tb$ value. Let $D_{ij}.TaTb$ represent the set of documents with a tag $Ta.Tb$ corresponding to the enconverted query graph g_{ij}. The next section describes how index based features are used to further rank each set of $D_{ij}.TaTb$ documents.

4.3 Use of Index based Features

Index based features are used to calculate a weight factor to prioritize the documents within each set $D_{ij}.TaTb$. The features used are position, frequency count, Named Entity(NE) tag and Multi-word(MW) tag of the term/concept. The feature weight is calculated as follows.

Index based feature Weight

$$W_I = \frac{P^i_{Weight} + F^i_{Count} + NE^i_{Weight} + MW^i_{Weight}}{\sum_{j=1..n}\left(P^j_{Weight} + F^j_{Count} + NE^j_{Weight} + MW^j_{Weight}\right)}$$

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Here i represents the single document weight and j represents the weight across the documents.

P^i_{Weight} represents the position weight of the concept. Position weight is computed based on where in the document the concept or term occurs

F^i_{Count} represents the frequency of occurrence of concepts in the document.

NE^i_{Weight} represents the Named Entity weight associated with the concepts in Q.

MW^i_{Weight} represents the Multi Word weight associated with the concepts in Q.

4.4 Computing Overall Ranking

The first step in computing the overall ranking is to merge all the $D_{ij}.TaTb$ documents corresponding to each gij of the Query Q. Those documents which occur in the maximum number of sets are ranked higher. The merged set of documents are then ranked based on $TaTb$ value, and each set $DTa.Tb$ is in turn ranked using normalized index based weight factor. δ is the normalized weight factor to differentiate between the complete CRC match,partial CR or C match. Here

$$\delta = \left\{ \begin{array}{ll} 0.5 & ifTa = 1 \\ 0.3 & ifTa = 2,3,4 \\ 0.2 & ifTa = 5,6,7 \\ & \end{array} \right\}$$

Thus, in this algorithm, the first level of ranking is obtained by Ta, in turn the set of documents corresponding to each Ta at the second level the documents are ranked according to Tb and then within the set, at level three these documents are ranked according to a normalized index based weight $\delta \times W^Q_{Ta.Tb}$.
Thus the conceptual searching and ranking algorithm considers degree of match,context of query, concept association and index based term,concept and position factors corresponding to sentences as well as documents for effective ranking.

5 PERFORMANCE EVALUATION

This system has been implemented for the Tourism domain and has been tested with a corpus of 33000 documents. For comparison purposes, a key word based search built by the CLIA (Cross Lingual Information Access) consortium [1] has been used with the same corpus. We have used MAP (Mean Average Precision) score for measuring the relevance of the documents retrieved, and Discounted cumulative gain for evaluating the ranking. MAP (Mean Average Precision) (Thom J et.al.,2007) is an arithmetic mean of average precisions over a set of queries used for evaluation. It calculates the average precision of every single query and then takes the mean value of all queries. The relevance judgment of each retrieved page is done based on the human judgment of the documents.

- If the resulting search documents not only contain the keyword but also other related information relevant to the user query, then the score will be 1.
- If the resulting search documents contain the user entered query terms but not describing the relevant information of the user entered query, then the score will be 0.
- A non responding URL is also given a score of 0.

A query set of 139 queries has been used for the evaluation. The MAP Score for 139 queries is given in the *Table*1 for the UNL systems. It can be seen that the MAP score of the UNL system is found to be around 0.72 even at the top 20 documents.The *Figure*.1 shows the MAP(Mean Average Precision) score at various top level results of UNL search system. The comparison of UNL search system with the keyword based search engine is given in *Table*.2 and the comparison chart is shown in *Figure*.2.

Table 1: MAP Score at various top level results for 139 queries of CLIA Advanced UNL Search System.

Relevance Judgement for	MAP Score
Top 5 documents	0.7625
Top 10 documents	0.7413
Top 20 documents	0.7243

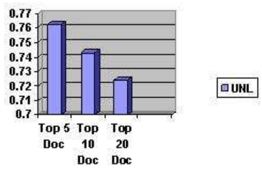

Figure 2: MAP Score of UNL.

Table 2: Comparison of CLIA Advanced UNL Search System with keyword based search systems

UNL	CLIA
0.7625	0.45

We have also compared our system with the popular search engine Google. We find that the MAP score is almost the same as that of the UNL search.

[1]A project funded by Ministry of Information Technology, New Delhi

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

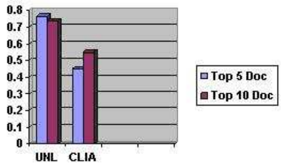

Figure 3: MAP Score of UNL.

However, the corpus used by Google is large and covers various domains, whereas our system is implemented only for the tourism domain. Thus, the list of documents is different. Hence a MAP score based comparison is not in order. Hence we have resorted to a controlled experiment in order to compare the performance of the UNL based search with that of Google. We have examined the results for concept related information, and find that there are a certain set of multiple-word queries for which the UNL system gives a better set of results in terms of relevance. This is a motivation to expand the system to other domains to perform a full-fledged comparison.

ஆக்ராவைச் சுற்றியுள்ள சுற்றுலாத்தலங்கள் (Agravai Chuttriyulla Chuttrulaaththalangal)- Tourist places near Agra
அரண்மனை உள்ள நகரம்(Aranmanai Ulla Nagaram)
தஞ்சாவூர்(Thanjavur), நெல்லை(Nellai) ,மாமல்லபுரம் (Mamallapuram) , ஊட்டி (Ooty), சிம்லா(Simla), காஞ்சி (Kanchi)
வைகை ஆறு உருவாகுமிடம்(Vaigai aaru uruvagumidam)-Origin of Vaigai river
மெரீனா பீச்(Marina beach)
ஆறு உருவாகுமிடம் (Aaru uruvagumidam)-Orgin of a River
தமிழ்நாட்டு உணவுகள் (Tamilnattu unavukal) -Foods in Tamilnadu
தமிழ்நாட்டின் ஜீவநதி (Tamilnattin Jeevanathi)
கேரளா பண்பாடு மற்றும் கலாச்சாரம்(Kerela panpadu mattrum kalaachchaaram) – Culture of Kerela
கந்தன் ஆலயம்(Kandhan Alayam)-Murgan Temple

Figure 4: List of Queries that yields best results than Google.

A list of queries in which our system ranks and retrieves better results than Google are: The computational complexity of the searching and ranking is also analyzed. The UNL indexer stores the indices in a Binary Search Tree. Therefore, the complexity of finding a node from the Binary Search Tree is O(logn). For sorting the rank results with respect to weight factor, insertion sort is used which has the time complexity O(logn). But the results are sorted immediately after retrieving, for CRC,CR and C respectively which further improves the search time. The average search time is 0.33 Seconds on a 8 GB RAM and 2.6GHz processor.

6 CONCLUSIONS AND FUTURE WORK

The index based query expansion to account for domain context is an important aspect of this work. The searching and ranking algorithm described in this work is based on three level ranking procedure, the degree of match level, nature of the concept association level and the index based feature prioritization level.

These levels of ranking help to fine tune searching and ranking in number of ways. The differentiation between actual query relation or other relation between CRC matches and the part of CRCs such as Source node or destination node of the CR and C sub graph in partial match helps to bring in more ranked categorization at the concept association level. This fine tuning of priority among the ranked documents is an important contribution of this work. Future work is to consider the role of differentiating UNL semantic constraints in the ranking machanism. The query expansion can also consider predicting relations between query concepts in multi word queries to build the UNL sub graphs. Additionally the categorization of certain relations as having higher priority in specific domain during ranking can also be studied.

REFERENCES

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117. Elsevier Science Publishers B. V.

Gao, M., Liu, C., and Chen, F. (2005). An ontology search engine based on semantic analysis. *Information Technology and Applications, International Conference on*, 1:256–259.

HAKIA (2009). A hakia search engine.

Khare, R. and Cutting, D. Nutch: A flexible and scalable open-source web search engine. Technical report.

Kraines, S. B., Guo, W., Kemper, B., and Nakamura, Y. (2006). Ekoss: A knowledge-user centered approach to knowledge sharing, discovery, and integration on the semantic web. In Cruz, I. F., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., and Aroyo, L., editors, *International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 833–846. Springer.

M, Z. and H, U. (1998). The universal networking language (unl) specification version 3.0 1998. In *Technical Report*.

Sensebot (2009). Sensebot search engine.

Subalalitha, Geetha, T. V., Ranjani, P., and Madhan, K. (2009). A concept based semantic indexing technique. International Conference on Web Intelligence Systems ICWIS09.

Surve, M., Singh, S., Kagathara, S., Dubey, S., Rane, G., Saraswati, J., Badodekar, S., Almeida, A., Nikam, R., Perez, C. G., and Group, A. (2004). Agro-explorer: A meaning based multilingual search engine. In *International Conference on Digital Libraries (ICDL*.

Tumer, D., Shah, M. A., and Bitirim, Y. (2009). An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia. In *Proceedings of the 2009 Fourth International Conference on Internet Monitoring and Protection*, pages 51–55.

UNDL (2009). Universal networking digital language foundation.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS