QUERY PROCESSING FOR ENTERPRISE SEARCH WITH WIKIPEDIA LINK STRUCTURE

Nihar Sharma and Vasudeva Varma
SIEL (LTRC), International Institute of Information Technology, Gachibowli, Hyderabad, India

Keywords: Query Expansion, Wikipedia link Graph, Thesaurus, Enterprise Search, Information Retrieval.

Abstract: We present a phrase based query expansion (QE) technique for enterprise search using a domain independent concept thesaurus constructed from Wikipedia link structure. Our approach analyzes article and category link information for deriving sets of related concepts for building up the thesaurus. In addition, we build a vocabulary set containing natural word order and usage which semantically represent concepts. We extract query-representational concepts from vocabulary set with a three layered approach. Concept Thesaurus then yields related concepts for expanding a query. Evaluation on TRECENT 2007 data shows an impressive 9 percent increase in recall for fifty queries. In addition to we also observed that our implementation improves precision at top k results by 0.7, 1, 6 and 9 percent for top 10, top 20, top 50 and top 100 search results respectively, thus demonstrating the promise that Wikipedia based thesaurus holds in domain specific search.

1 INTRODUCTION

Enterprise Search (ES) is a critical performance factor for an organization in today's information age. Intelligent algorithms have been designed to retrieve both structured (relational databases) and unstructured information (natural text) for businesses. The paradigm has shifted towards semantic search for both web and ES (Demartini, 2007). However, retrieval of unstructured information in an enterprise environment is a non trivial task as the relevance is the prime objective unlike its counterpart which focuses on speed.

We present our approach of harnessing semantics from an external domain independent knowledge resource for improving search process within enterprise environment while keeping focus on text retrieval. Our idea concentrates on Query Expansion (QE) aspect of search cycle. The key novelties in our approach are use of Wikipedia as the knowledge resource and treatment of multiword queries as singular conceptual entities for extracting synonymous concepts for expansion. Going a step beyond conventional thesaurus based QE techniques, which perceives a search query as collection of key words and each word was regarded as a separate concept.

The subsequent sections put forward an in depth analysis of our approach and its evaluation. In section 2, we discuss the motivating research findings behind our approach and we also look into the related work. Section 3 focuses on our research methodology and presents our implementation of an efficient QE module based on concept vocabulary and concept thesaurus derived from Wikipedia link graph, for an ES system. In section 4, we validate the performance of our model using TREC 2007 Enterprise Search Track data. Sections 5 and 6 present future paths for our conceived ideas and the conclusion respectively.

2 MOTIVATION AND RELATED WORK

Our motivation mainly arises from work done on role based personalization in enterprise search (Varma, 2009). The system used word co-occurrence thesaurus and WordNet for QE. We noticed the concept coverage of WordNet like manual thesaurus was mostly limited to single word concepts and it severely lacked named entities. This restricted the outlook of the QE module of a multiword search query to a bag of words perspective.

Sharma N. and Varma V..
QUERY PROCESSING FOR ENTERPRISE SEARCH WITH WIKIPEDIA LINK STRUCTURE .
DOI: 10.5220/0003093702430249
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 243-249
ISBN: 978-989-8425-28-7

Recognition of Wikipedia as a knowledge base among research groups across the globe was another inspiration for us to proceed in this direction. Medelyan et al. (Medelyan, 2009) give a comprehensive study of ongoing research work and software development in the area of Wikipedia semantic analysis. Their findings reveal a large number research groups involved with Wikipedia thus indicating its increasing popularity. A number of developments with Wikipedia content and link structure analysis are going on, which are related to our work. Ito et al. in (Ito, 2008) present their approach on constructing a thesaurus using article link co-occurrence where they relate articles with more than one different path between them, within a certain distance in the link graph. Nakayama et al. present another approach for creating thesaurus from large scale web dictionaries like Wikipedia using a concept of path frequencies and inverse backward link frequencies in a directed link graph of concepts (Nakayama, 2007). Milne et al. introduce a search application using Wikipedia powered assistance module that interactively builds queries and ranks results (Milne, 2007). However, it does depend on active user involvement during the search cycle. The idea of using Wikipedia as a domain independent yet comprehensive knowledge base is consolidated by the findings of Milne et al. (Milne, 2006). They give a comprehensive case study of Wikipedia concept coverage of domain specific topics.

3 QUERY EXPANSION WITH WIKIPEDIA CONCEPT THESAURUS

We present a QE technique that involves an external domain independent knowledge resource, the English Wikipedia. Our approach builds a Concept Thesaurus (CT) by analyzing inter-article and article-category link graphs. Our work mainly focuses on Wikipedia, however the algorithms we are proposing can work on any document collection having a substantial number of interlinks. We also propose a query to concept mapping mechanism which can extract representative concepts for a given query. Using query-concept mapping and CT, we put forward a QE module which adds related concepts to a query. Figure 1 and 2 sum up the process and following sub-sections explain it in detail.

3.1 Wikipedia Composition and Data

A Wikipedia page can be an article, a redirect, category, talk, special et cetera. Majority of these pages are informative articles describing more than 2.5 million topics from almost every knowledge domain. In addition to articles, we are also interested in Redirect and Category pages. Articles are richly linked to other articles (inter-article links) from within their text. Articles are categorized by linking them to respective category pages by article-category links (English Wikipedia has around 450.000 categories).

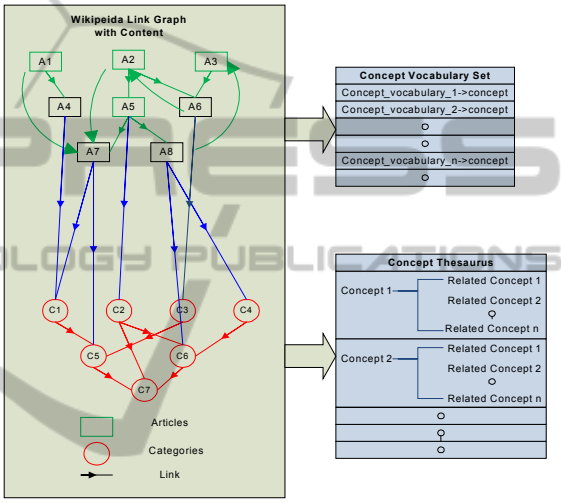

Figure 1: Wikipedia Link Structure and Creation of Vocabulary set and Thesaurus.

3.2 Concept Representation and Concept Vocabulary

The articles in Wikipedia have a title and text description explaining a concept. The concept name is represented by the Article title. The article title is an un-ambiguous precise word or phrase that best represents the concept explained by the article. These titles are good candidates to be used as concept entities in a thesaurus of Wikipedia concepts.

In addition to article names, redirects and anchor texts can be additional representations of concepts. Redirects are alternative concept names that could be used as a title of an article but instead they just link to it. For example, an article on the First World War in Wikipedia has the title 'World War I' and few redirects that connect to it have the titles 'WW1', 'First World War', 'The Great War' etc.

Anchor text also represents a concept. Usually, the anchor texts are part of natural sentences and

cover various forms and inflections of concept representational words. To clarify, inflections of the word molecule, like 'molecules' and 'molecular' are anchor texts in Wikipedia articles connecting to 'Molecule' article. Further, the similar meaning of structurally different phrases is also captured with anchor text representation (for example, anchor texts "possessed by the devil", "Demonic possession" and "Control over human form" link to same article and thus have similar meanings), not to mention the inclusion of acronyms (anchor text 'WWW' links to the article 'World Wide Web').

Wikipedia concept vocabulary is stored in a dictionary based data-structure along with the target articles it represents. Article Titles and Redirect Titles implicitly have one to one relationship with concepts but same Anchor Text can have more than one target concepts. For resolving anchor to multi-concept relation, we employ a simple measure to establish one target article per anchor text in the vocabulary set. For all multi-concept anchors, an article that is most number of times linked by an anchor text in all Wikipedia inter-article links is ascertained as its final target concept. It may seem a trivial way to resolve the target conflicts for anchor text, but this approach saves us significant computational expense which otherwise would have occurred with a more adaptive but complex approach like TF-IDF.

3.3 Mining Concept Thesaurus from Wikipedia Links

We condense the knowledge of Wikipedia article content and links into a Concept Thesaurus (CT), which is not just a set of synonyms, but a set capturing all logical relations between concepts (relations like ‘water’ to ‘ocean’). Analyzing article text, shows us that the text contains many such logically related concept phrases. Many of these phrases are turned into hyperlinks that connect to their own articles. We capitalize on these links and narrow down our focus to mutually linked articles, which we conceive as related concepts, as the relation is validated by two way link created by human intelligence. We also pay attention to linked concepts which share a common domain, in other words belong to common categories.

3.3.1 Cross Link Analysis

First step involves examining a concept and all the inter-article links from the article explaining a concept (let's call this the article under examination 'A'; we would use 'A' for representing the concept as well). Among the articles linked from 'A', the ones which have a link back to 'A' represent mutually related concepts.

3.3.2 Link Co-category Analysis

In the next step, we look into one-way links from 'A'. This step also involves the category pages linked to 'A' (categories of 'A'). An article which has a link from 'A' and belongs to one of the categories of 'A', represents a related concept. An important point to note here, a small but significant number of Wikipedia categories are related to article status and do not indicate a concept domain (categories like ‘stub article’, ‘articles to be deleted’ etc.). We have taken care to implement a filter that weeds out analysis of such categories.

3.3.3 Relation Specific (RS) Score

After determining related concepts, we resolve how closely related two articles are, based on the study of overall Wikipedia link structure as well as the text analysis of articles related to 'A'. All candidate concepts related to 'A' are given the RS score. If an article 'AR' is related to 'A' (hence the concepts represented by A and AR), then

RS ('A', 'AR') = count (title ('A'), text ('AR')) / inLink ('AR')

Where, RS ( ) is the RS score of relation between ‘A’ and ‘AR’, count (str, tx) is number of times name of string ‘str’ appeared in text ‘tx’, and ‘inLink’ (article) is the number of backward links to the article. Ding et al. (Ding, 2005) explain a backward link as a simple relation between articles ‘A1’ and ‘A2’, if ‘A2’ has a hyperlink targeting ‘A1’.

Articles having a relatively higher number of other articles linking to them thus, represent more generic concepts; and will be given a low RS score from above formula. A complete analysis of Wikipedia article link structure, category network and article text using the above three steps yields a CT of related concepts with respective RS scores.

Algorithm 1 implements the CT extraction process. Lines 7 to 14 extract cross linked articles for every concept. The link co-category analysis is carried out by lines 15 to 20. Lines 21 to 29 determine the RS scores for all related concepts. The output of this algorithm is the final CT.

Algorithm 1: Wikipedia CT extraction.

```
Input: AL    # Wikipedia article list with id, title,
         content information
1: CT ← ∅
2: for all article in AL do
3:   categorySet ← getCategories (article)
4:   linkSet ← getLinkedArticles (article)
5:   vocabList ← getConceptVacab(article)
6:   for all linkedArticle in linkSet do
7:      blnRelated ← false
8:      linkSetTemp ←
        getLinkedArticles (linkedArticle)
9:      for all link in linkSetTemp do
10:       if link == article then
11:         blnRelated ← true
12:         break for
13:       end if
14:     end for
15:     if not blnRelated then
16:       categorySetTemp ←
          getCategories (linkedArticle)
17:       if commonCategoryExists (categorySet,
                categorySetTemp) then
18:         blnRelated ← true
19:       end if
20:     end if
21:     if blnRelated then
22:       conceptCount ← 0
23:       for all vocab in vocabList do
24:           conceptCount ←
                conceptCount +
                count (vocab, text(linkedArticle))
25:       end for
26:       inFrequency ←
            countIncomingLinks (linkedArticle)
27:       score ← conceptCount / inFrequency
28:       addToThesaurus (CT, article,
                linkedArticle, score)
29:     end if
30:   end for
31: end for
```

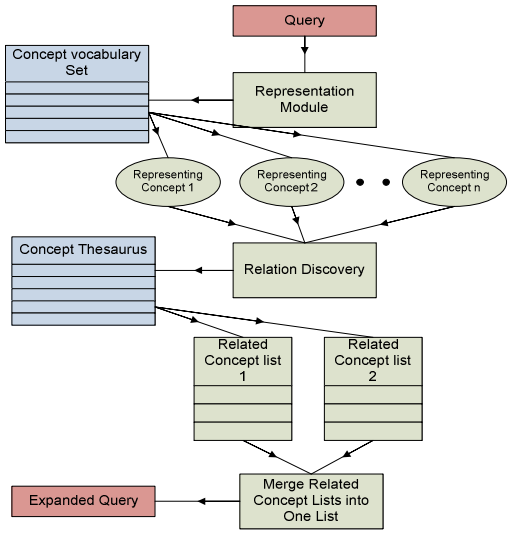

Figure 2: Concept Representation and Query Expansion Module.

3.4 Wikipedia Concept Representation of Search Query

In order to use CT, we need a set of concepts that can best represent a query. User queries frequently lack clarity and the word structure (order of words, inflections) where as much of the concept vocabulary is a set of standard article names (precise word order), hence steps beyond simple word matching are required for mapping a query to concepts. Moreover, a query may require a combination of concepts to represent its scope. Another significant requirement of multi-word query mapping is to capture its sense in its entirety and not just as collection of individual key terms, in other word capturing the phrase behaviour of a search query.

We represent queries with best N (number of representations required) Wikipedia Concepts through three levels of concept discovery. We exploit Concept Vocabulary and article content for determining the concepts that depict search query. Along with maintaining a vocabulary-target list we build a TF-IDF based index on vocabulary phrases (vocabulary index) and on article content text (content index).

Algorithm 2 implements the above process. First level (line 7, Algorithm 2) involves string comparison of query and concept vocabulary entries. Figure 2 (first half of the figure, above relation discovery part) represents this process abstractly. An exact match, if available, fetches the best representation of query and is stored as level one concept (EMVT, Algorithm 2). Second level involves looking through partial matches of vocabulary phrases based on cosine similarity scores between vocabulary index and query (line 8, Algorithm 2). Concept names for top N matches are stored in second level match list (CMVT, Algorithm 2). Third level requires fetching top N matches while comparing query with content index (line 9, Algorithm 2). Concepts represented by the fetched content pages are stored in third level list (CMCT, Algorithm 2).

Subsequent to the three levels of Wikipedia concept discovery, we start building the concept representation list (lines 10 to 32, Algorithm 2). We prioritize the fetched concepts based on their levels. First entry in representation list (RL) is EMVT (line 11, Algorithm 2) if it is not empty. We consequently use CMVT entries (lines 14 to 22, Algorithm 2) to append RL with a cosine similarity threshold (T, Algorithm 2) check between query and concept names. After CMVT is exhausted, CMCT elements

Algorithm 2: Determination of query representational Wikipedia Concepts.

```
Input: CV, CVI, CCI, Q, N
# Concept Vocabulary, concept Vocabulary Index,
# Concept Content Index and Query respectively
1: RL ← ∅          # List of representation concepts
2: EMVT ← “”   # String for exact query vocabulary match
3: CMVT ← ∅   # List of concepts for query vocabulary match on
cosine similarity
4: CMCT ← ∅   # List of concepts or query content match on
cosine similarity
5: T              #integer, cosine similarity Threshold, (empirically
determined)
6: count ← 0
7: EMVT ← getExactMatchConcept (CV, Q)
8: CMVT ← getTopMatch (CVI, Q, N)
9: CMCT ← getTopMatch (CCI, Q, N)
10: if EMVT != “ then
11:     addToList (EMVT, i)
12:     count ← count + 1
13: end if
14: for all title in (CMVT) do
15:     if cosSimilarity (Q, title) > T then
16:         addToList (title, RL)
17:         count ← count + 1
18:         if count > N then
19:             break for
20:         end if
21:     end if
22: end for
23: QC ← removeStopWords (Q)
24: for all title in (CMCT) do
25:     if partialTermMatch (QC, title) then
26:         addToList (title, RL)
27:         count ← count + 1
28:         if count > N then
29:             break for
30:         end if
31:     end if
32: end for
33: return RL
```

are used to append RL (lines 23 to 32, Algorithm 2) and a threshold check similar to CMVT check is employed. The process of appending RL is stopped as soon as its size reaches N.

3.5 Query Expansion

Once the Wikipedia concept representation is established for a user query, we retrieve a list of N most related concepts for every concept representing the query from CT. These related concepts include semantically related senses to the query. Next we merge these lists. Common concepts across these lists have their respective RS scores added in order to increase the importance of a concept that appears more frequently. Final list is sorted in a decreasing order of RS scores and top N concepts are added to Q. Repetitions and stop words are removed from the reconstructed query. This process is represented in Figure 2, lower half.

4 EVALUATION AND RESULTS

In 2007, the CSIRO Enterprise Research Collection corpus was introduced as the Enterprise Search evaluation data for TREC (TRECENT, 2007). Both our baseline and CT based QE setups are evaluated on plain text search on TRECENT and we did not look into internal link structure of the evaluation data (because Enterprise Data in general, lacks the strong link structure which is present in web domain (Mukherjee, 2004)).

4.1 Baseline Setup

For creating the test set, 50 queries were formulated from the 50 topics provided, where every query was a Boolean OR combination of its individual terms present in the query.

4.2 Query Expansion Setup and Evaluation

Wikipedia link structure database dumps were loaded on MySQL database and article content was stored and indexed using Apache Lucene indexer. All the modules for extracting concept vocabulary, concept relations and query representations were coded in java. The concept vocabulary set as well as CT were also tabulated in MySQL. A concept vocabulary index for partial query phrase matching was also created.

We configured Query Representation module to extract two best concepts which it generated on the run using vocabulary database and Lucene index. We tweaked the QE module to add 2 most related concepts to the query for the first representing concept and 1 related concept for second one. Following are couple of sample expansions created by our CT based QE module, are:

1. **Query**: timber treatment
   Reconstructed Query: timber treatment + Wood preservation + Chromated copper arsenatem + Creosote
2. **Query**: cancer risk
   Reconstructed Query: cancer risk + Cervical cancer + Human papillomavirus + Pap test

We recorded Precision for top 10, 20 50 and 100 retrieved results and overall Recall measures of both setups to compare their performance. Table 1 gives the respective observations of these measures and shows an improvement of 0.7%, 1%, 6%, 9% for precision at top10, top 20, top 50 and top 100 search results respectively, averaged over 50 queries. The overall Recall has significantly improved by an impressive 9%.

Table 1: The evaluation results for Wikipedia CT based Query Expansion module.

	Baseline	CT based QE	Percent Increase
Number of Queries	50	50	-
Average number of Results Retrieved	15901	28182	77.2%
Recall	0.834	0.910	**+9.1%**
Precision (over all)	0.074	0.047	-
Precision at top 10	0.693	0.698	**+0.72%**
Precision at top 20	0.593	0.599	**+1.01%**
Precision at top 50	0.457	0.485	**+6.12%**
Precision at top 100	0.377	0.412	**+9.2%**

The First striking observation from the results is the impressive 9% increase in the recall figures. Moreover, recall has reached above 0.9 for most queries thus indicating the effectiveness of concept coverage of QE module. However, increase in recall is consistent with most QE techniques. It is the non degradation of precision at top 10 results and improvements for top 20, top 50 and top 100 results that impress the most. Thus if we compare the average number of documents retrieved per query, we see that in spite of a relatively much larger number of retrieval for QE module, relevant results are not pushed back. Our experiments only involved query level adjustments and a combination of semantics based indexing and re-ranking techniques can further improve the precision of search results.

5 DISCUSSION AND FUTURE WORK

The use of an external knowledge resource for QE has been a tried and tested area in both web and Enterprise Search. WordNet is one such popular resource. With more than 2 million English articles, Wikipedia knowledge in terms of both number of concepts and Named Entities far exceeds any available dictionary or thesaurus in electronic format for English language.

We are working on extracting sub-graphs of the Wikipedia link structure, which will contain the corpus specific concepts. Another area we are aiming to work on is document term enrichment with concept vocabulary. Using techniques similar to concept representation we plan to add standard concept names into document. Our QE module introduces terms which are concept names. The syntactic structure of these terms might be different from semantically similar terms in the corpus, thus such document may not to be detected in a TF-IDF based search. Our idea will introduce the same vocabulary in the documents with which we are reconstructing a query.

6 CONCLUSIONS

We have put forward our technique to extract a CT by focussing on the inter-article link graph of Wikipedia as well as a way to map any search query to these concepts by treating it as semantic unit instead of a bag of words. Our aim is to make a generic ES Application which should be independent of enterprise structure for reasonable functioning without losing the power of customization and should require a minimal human intervention for configuration.

REFERENCES

Demartini G., 2007. Leveraging semantic technologies for enterprise search. In *Proceedings of the ACM first Ph.D. workshop in CIKM, PIKM*. ACM press.

Mukherjee R., Mao J., 2004. Enterprise Search: Tough Stuff. In *Queue, Volume 2, Issue 2*. ACM press.

Ding C., 2005. Probabilistic model for Latent Semantic Indexing: Research Articles. In *Journal of the American Society for Information Science and Technology, Volume 56 Issue 6*. ASIS&T.

Ricardo A. Baeza-Yates , Ribeiro-Neto B., 1996. *Modern Information Retrieval*. Addison-Wesley, Longman Publishing Co., Inc., Boston, MA.

Xu J, Croft W. B., 1996. Query expansion using local and global document analysis. *In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM press.

Medelyan O., Milne D., Legg C., Witten I. H., 2009. Mining meaning from Wikipedia. In *International Journal of Human-Computer Studies, Volume 67, Issue 9*. Elsevier.

Milne D., Witten I. H., Nichols D. M., 2007. A knowledge-based search engine powered by wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM), Lisbon, Portugal*. ACM press.

Milne D., Medelyan O., Witten I. H., 2006. Mining Domain-Specific Thesauri from Wikipedia: A Case Study. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE.

Masahiro Ito, Kotaro Nakayama, Takahiro Hara, Shojiro Nishio, 2008. Association thesaurus construction methods based on link co-occurrence analysis for wikipedia. *In Proceeding of the 17th ACM conference on Information and knowledge management (CIKM '08)*. ACM press.

Kotaro Nakayama, Takahiro Hara, Shojiro Nishio, 2007. A Thesaurus Construction Method from Large ScaleWeb Dictionaries. In *Proceeding of the 21st International Conference on Advanced Networking and Applications*. IEEE.

Varma V., Pingali P, Sharma N., 2009. Role Based Personalization in Enterprise Search. *4th Indian International Conference on Artificial Intelligence, Special Session on Web 2.0 and Natural Language Engineering Tasks, Bangalore, India*.

TREC Enterprise Track: TRECENT 2007. http:// trec.nist.gov/tracks.html