ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)
On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Effective E-Learning Recommender System Using Query Expansion Technique in Information Retrieval

Anitha.V[1], Ravichandran. M[2]

ME, Department Of Computer Science & Engg., Dept. of IT, Arunai Engg. College, Tiruvannamalai, Tamilnadu, India

ME, Department Of Computer Science & Engg., Dept. of IT, Arunai Engg. College, Tiruvannamalai, Tamilnadu, India

ABSTRACT— In an information retrieval system users cannot accurately give their queries for retrieving a particular context. So the term mismatch problem occurs (i.e.) a user query for Information Retrieval applications do not contain the appropriate terms as actually intended by the user. The fundamental issue in the Information Retrieval System is term mismatch or word mismatch problem, as we already know that the effective way to handle the problem is query expansion technique. Query expansion adds related term to original query, which provides more information about the user needs. This paper implements a well-known Global Query Expansion tool namely WORDNET for expanding the query in any given Information Retrieval systems. Global Query Expansion comprises of Similarity thesaurus and Statistical thesaurus. This paper includes similarity thesaurus. Global similarity thesaurus has to be computed only once and can be updated incrementally. Also this paper incorporates a representation model named bag-of-words which makes this technique effective, simple and convenient. This paper calculates the similarity values, so that the result will be improved in its accuracy and performance. The result shows flexible and simple execution by reducing the run-time computational overhead.

INDEX TERMS—Information Retrieval, Query Expansion, Word- net, Bag-of-words

I. INTRODUCTION

In a real scenario, however, the query submitted by the user is not the only information about the information needs of the user [1]. Users typically formulate very short queries [2] and they may use different words than the source document to describe the same concept, giving rise to the term mismatch problem [3]. Query Expansion is the term given when a search engine adding search terms to a user's weighted search. The goal is to improve precision and/or recall. Example: User Query: "car", Expanded Query: "car, cars, automobile, automobiles, auto" etc...

Query expansion can either be manual or automatic.

1) Interactive Query expansion (IQE) is an example of manual technique (approaches based on feedback infor- mation from users).
2) Automatic Query Expansion (AQE) is an example of Automatic technique that offers a specific advantage over manual because they do not require any user effort. Automatic query expansion (AQE) has been suggested as a technique for dealing with the issue of word mismatch or term mismatch in Information Retrieval. AQE is broadly categorized into two types, global and local [4].

Fig. 1. Typical Architecture Diagram

II. RELATED WORK

Automatic query expansion has been considered for last four decades. The requirement for Automatic Query Expansion was known because of the intrinsic limitation of the conventional information retrieval systems. A user may communicate query terms which do not match the terms appearing in the relevant documents, giving rise to the vocabulary problem [3]. The vocabulary problem is also termed a "word mismatch" prob- lem or "term mismatch" problem. The problem is particularly severe with short queries which are becoming increasingly common in retrieval applications [5]. Query expansion is an efficient technique normally used to add useful terms to the user query. This section gives a systematic analysis of a variety of query expansion techniques. We will describe global query expansion techniques that do not make use of the word ordering information for calculating term correlations. Then, a brief review of techniques that use bag-of-words for indexing purpose to obtain term correlations will be given. Next, we will discuss the WUC palmer technique to measure the similarity values, followed by the local techniques for query expansion.

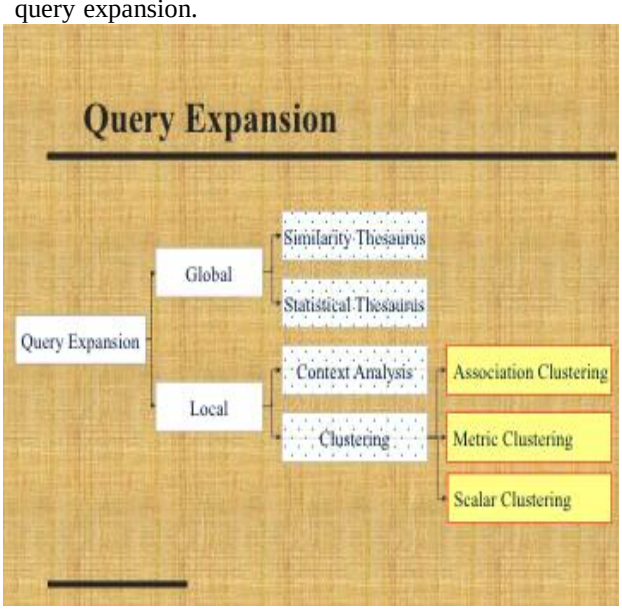

Fig.2.Query Expansion

III. GLOBAL QUERY EXPANSION

A Global technique based on global information derived from the document collection. There are two modern variants based on a thesaurus-like structure built using all documents in collection.

1) Query Expansion based on a Similarity Thesaurus.
2) Query Expansion based on a Statistical Thesaurus.

A. Thesaurus

A thesaurus provides information on synonyms and semantically related words and phrases.

Example:
physician
syn: ||croaker, doc, doctor, MD, medical, mediciner, medico,
||sawbones rel: medic, general practitioner, surgeon.

B. similarity thesaurus

The similarity thesaurus is based on term to term re-lationships rather than on a matrix of co-occurrence. This relationship is not derived directly from co-occurrence of terms inside documents. They are obtained by considering that the terms are concepts in a concept space. In this concept space, each term is indexed by the documents in which it appears. Terms assume the original role of documents while documents are interpreted as indexing elements. Add synonyms in the same synset.

Query expansion based on similarity thesaurus

- Represent the query in the concept space used for repre- sentation of the index terms
- Based on the global similarity thesaurus, compute a similarity sim(q,kv) between each term kv correlated to the query terms and the whole query q.
- Expand the query with the top r ranked terms according to sim(q,kv).

C.Statistical Thesaurus

Existing human-developed thesauri are not easily available in all languages. Human thesuari are limited in the type and range of synonymy and semantic relations they represent. Semantically related terms can be discovered from statistical analysis of corpora.

Query expansion based on statistical thesaurus

- Global thesaurus is composed of classes which group correlated terms in the context of the whole collection
- Such correlated terms can then be used to expand the original user query
- This terms must be low frequency terms
- However, it is difficult to cluster low frequency terms
- To circumvent this problem, we cluster documents into classes instead and use the low frequency terms in these documents to define our thesaurus classes.

IV. LOCAL QUERY EXPANSION

Local technique based on information derived from the set of documents initially retrieved (called the

local set of documents).Local techniques works on the principle of pseudo relevance feedback (PRF) [6].

In local technique two types can be used

1) Query expansion through local clustering (Local Feed- back)
2) Query expansion through local context analysis

A. Local feedback strategies

Based on expanding the query with terms correlated to the query terms. Such terms are those present in local clusters built from the local document set. The types of clusters are Association clusters, Metric clusters and Scalar clusters.

B. Local context analysis

Uses noun groups instead of keywords as concepts for query expansion extracted from top retrieved documents .Uses document passages for determining co-occurrence.

V. MOTIVATION

Information retrieval is broadly concerned with the problem of organizing collections of documents to support various In- formation requests by users. Information retrieval is the activ- ity of obtaining an information resource relevant to needed in- formation from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications. We use the system relevance criteria for the theoretical development of the proposed query representation (QR) framework. Empirical evidence from the user relevance criteria is used to further justify the theoretical derivation. Indexing terms are used to give a Òbag-of wordsÓ representation to both the queries and documents, consistent with the practice of IR systems.

VI. METHODOLOGY

A. Bag Of Words

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregard- ing grammar and even word order but keeping multiplicity. Recently, the bag-of-words model has also been used for computer vision. The bag-of-words model is commonly used in methods of document classification , where the (frequency of) occurrence of each word is used as a feature for training a classifier.

B. Intrinsic vs. extrinsic evaluation

Intrinsic evaluation considers an isolated NLP system and characterizes its performance mainly with respect to a gold standard result, pre-defined by the evaluators. Extrinsic evalu- ation, also called evaluation in use considers the NLP system in a more complex setting, either as an embedded system or serving a precise function for a human user. The extrinsic performance of the system is then characterized in terms of its utility with respect to the overall task of the complex system or the human user. For example, consider a syntactic parser that is based on the output of some new part of speech (POS) tagger. An intrinsic evaluation would run the POS tagger on some labeled data, and compare the system output of the POS tagger to the gold standard (correct) output. An extrinsic evaluation would run the parser with some other POS tagger, and then with the new POS tagger, and compare the parsing accuracy.

C. Black-box vs. glass-box evaluation

Black-box evaluation requires one to run an NLP system on a given data set and to measure a number of parameters related to the quality of the process (speed, reliability, resource consumption) and, most importantly, to the quality of the result (e.g. the accuracy of data annotation or the fidelity of a trans- lation). Glass-box evaluation looks at the design of the system, the algorithms that are implemented, the linguistic resources it uses (e.g. vocabulary size), etc. Given the complexity of NLP problems, it is often difficult to predict performance only on the basis of glass-box evaluation, but this type of evaluation is more informative with respect to error analysis or future developments of a system.

D. Automatic vs. Manual evaluation

In many cases, automatic procedures can be defined to evaluate an NLP system by comparing its output with the gold standard (or desired) one. Although the cost of producing the gold standard can be quite high, automatic evaluation can be repeated as often as needed without much additional costs (on the same input data). However, for many NLP problems, the definition of a gold standard is a complex task, and can prove impossible when inter-annotator agreement is insuffi- cient. Manual evaluation is performed by human judges, which are instructed to estimate the quality of a system, or most often of a sample of its output, based on a number of criteria. Although, thanks to their linguistic competence, human judges can be considered as the reference for a number of language processing tasks, there is also considerable variation across their ratings. This is why automatic evaluation is sometimes referred to as objective evaluation, while the human kind appears to be more “subjective”.

VII. WORDNET TOOL

WordNet is a lexical database for English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. The database and software tools have been released under a BSD style license and can be downloaded and used freely. The database can also be browsed online.

WordNet was created at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George Armitage Miller. Wordnet is still maintained by the Cognitive Science Laboratory [7].

WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. It does not include prepositions, determiners etc. Every synset contains a group of synonymous words or collocations (a collocation is a sequence of words that go together to form a specific meaning, such as "car pool"); different senses of a word are in different synsets. The meaning of the synsets is further clarified with short defining glosses (Definitions and/or example sentences). A typical example synset with gloss is:

good, right, ripe Ð (most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes")

Most synonym sets are connected to other synsets via a number of semantic relations. These relations vary based on the type of word, and include:

• WordNet Synset Relationships:

– Antonym: front - back
– Attribute: benevolence - good (noun to adjective)
– Pertainym: alphabetical - alphabet (adjective to noun)
– Similar: unquestioning - absolute
– Cause: kill - die
– Entailment: breathe - inhale
– Holonym: chapter - text (part-of)
– Meronym: computer - cpu (whole-of)
– Hyponym: tree - plant (specialization)
– Hypernym: fruit - apple (generalization)

• Role of Automated Reasoning:

– look up a word and see definitions and examples of its word senses
– follow the primitive relations from a word sense to other word senses
– see statistics on words, word senses, synonymy and polysemy

• WUcPalmer:

– This measure calculates relatedness by considering the depths of the two synsets in the WordNet tax- onomies, along with the depth of the LCS WUP(s1, s2) = 2*dLCS.depth / (min_dlcs in dLCS(s1.depth - dlcs.depth)) + min_dlcs in dLCS(s2.depth - dlcs.depth)), where dLCS(s1, s2) = argmax_lcs in LCS(s1, s2)(lcs.depth).

– Parameters:
– min score = 0.0

TABLE I

DIFFERENTVERSIONOFQUERIES

	Full	SmryCon	Summary
Mean Number	52.54	29.22	11.02
Mean Ration	0.36	0.7	1.7

– max score = 1.0
– error score = -1.0
– acceptable pos pairs = [[’n’, ’n’], [’v’, ’v’]]
– use all senses = true
– use root node = true

VIII. EXPERIMENTS

Experiments were performed to empirically justify the proposed query representation model as a query expansion method using WORDNET tool. This paper examines the utility of lexical query expansion in the large, diverse TREC collection. Concepts are represented by WordNet synonym sets [8].

Algorithm 1 Procedure to automatically select synonym sets to expand

```
1: procedure SYNSET SELECTION(query)
2:     for all word in query do
3:         if word not already expanded
4: and document frequency of word < N) then
5: expand all synsets containing w producing kin list of w
6:         end if
7:     end for
8:     for all relative in set of kin lists do
9:         if relative occurs in more than 1 list then
10: add relative to query vector
11:         end if
12:     end for
13: end procedure
```

Table I compares the lengths of the different query vectors. The table contains the mean number of original terms and the mean ratio of additional terms to original terms for each of the different versions of queries: derived from full topic (Full), derived from Summary

Statement plus Concepts (SmryCon), and derived from Summary Statement (Summary) only.

a) Identifying Dataset:

- In this dataset, the user can give their queries as inputs in this module.
- The given inputs must be set of words.
- Searches can be based on metadata or on full-text index- ing.
- Input queries are needed to be relevant to the dataset.

b) Comparison with dataset:

- The user can give the inputs as words.
- words can be separately taken.
- Searches can be based on metadata or on full-text index- ing.

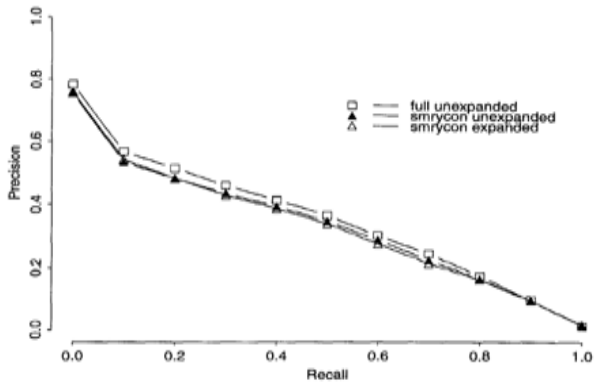

Fig. 3. Effectiveness of queries derived from summary statement and Concept fields.

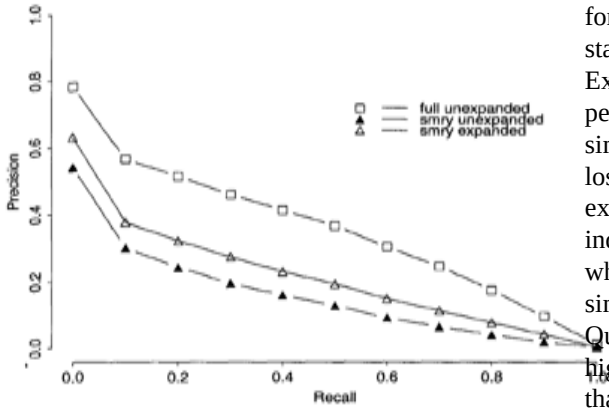

Fig. 4. Effectiveness of queries derived from only summary statement.

- Every word can be compared with dataset which has been already in our system.

c) Filtering the Mismatched term:

- The user inputs can be taken and then compared with the dataset.
- The matched words are going to next process.
- The mismatched words are filtered by the filters.
- Finally, filtered data's are going to next process.

d) Expanding the query:

- The mismatched words are collected and treated carefully.
- Then the word is expanded by means of Synset using WORDNET.
- After updating the bag of words, words can move to comparing.
- Finally it is added to the dataset.

e) Result Set:

- Words can move to comparing. After comparing and filtering the final result set should be ready, based on the weight of the word. The final result set is the original data that is related to the dataset.

- User can get the final result set.

IX. CONCLUSION

This paper performs a rigorous analysis for the desired query representation using WORDNET. The results show that Query representation gives statistically significant improve- ment over various datasets from TREC. The proposed method does not offer any extra computational burden than the other query expansion technique. Also the model representation bag- of-words makes the effort simple and convenient. This paper calculates the similarity value by means of similarity thesaurus of global query expansion technique. We go for similarity thesaurus since; query expansion based on statistical thesaurus need well defined parameters. Expansion of queries with related terms can improve performance, particularly recall. However, must select similar terms very carefully to avoid problems, such as loss of precision. The computation is ex- pensive but it is executed only once. The proposed model uses the indexing weights in the generalized retrieval framework, which can be directly used for the computing the similarity of a document with respect to user query. Query expansion based on similarity thesaurus may use high term frequency to expand the query. This illustrates that the proposed model can also be used as a framework to propose or validate a new query expansion method.

REFERENCES

[1] D. Downey, S. Dumais, D. Liebling, and E. Horvitz, "Understanding the relationship between searchers, queries and information goals," in 17th ACM conference on Information and Knowledge Management, 2008, pp. 449–458.

[2] Real Life, Real Users and Real Needs: A Study and Analysis of User Queries on the Web, ser. Information and Processing Management, vol.36 No. 2, 2000.

[3] The Vocabulary Problem in Human-System Communication, ser. Commu- nications, vol. 54 No. 8, ACM, 1987.

[4] Improving the Effectiveness of Information Retrieval with Local Context Analysis, ser. Information Systems, vol. 18 No. 1, ACM, 2000.
[5] Mining Longitudinal Web Queries: Trends and Patterns, ser. Information Science and Technology, vol. 54 No. 8, 2003.
[6] M. Okabe, K. Umemura, and S. Yamada, "Query expansion with a minimum user feedback by transductive learning," in Human Language Technology and Empirical Methods in Natural Language Processing (HLT), 2005, pp. 963–970.
[7] WordNet: An Online Lexical Database, vol. 3(4), 1990.
[8] E. M. Voorhees, "Query expansion using lexical-semantic relations," Siemens Corporate Research Inc., Tech. Rep., 1994.