International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-3 Issue-3, February 2014

An Efficient Converse Mapping Technique in Modern Information Retrieval

M. Ramya, E. Thenmozhi

Abstract: ***A traditional forward dictionary maps from words to their definitions. Unlike a regular dictionary, the reverse dictionary performs converse mapping that maps from definitions to words. The user input phrase describes the definition of desired concept, and returns an efficient candidate word that satisfies the input phrase. The reverse dictionary addresses the widespread problem of knowing the meaning of a word, but unable to recall the appropriate word on demand. The converse mapping technique finds the exact result for the user entered keyword by comparing the partitioned input with dictionary database. The partitioning method increases the overall scalability and distributes the data across multiple threads. The efficiency of the reverse dictionary can be improved by reducing the set of definitions in the comparison process. The query expansion technique is used to improve the potential of reverse dictionary and increases the probability of identifying relevant definition. The approaches of reverse mapping provide significant improvements in the performance scale. The converse mapping technique in modern information retrieval extracts the best matched result without sacrificing the quality of the solution.***

Keywords: ***Dictionaries, thesauruses, search process, web-based services***

I. INTRODUCTION

The converse mapping technique that maps from sense phrase to word phrase provide significant methods for creating and implementing an online reverse dictionary (RD). A forward dictionary takes the word phrase as input and returns the description of desired concept as sense phrase. For example, a forward dictionary informs the user that the meaning of the word “abseil” is “descent down a rock using rope”. As opposed to a regular forward dictionary, the reverse dictionary takes the description of desired concept as input and provide words that matches the entered definition phrase. For example, the reverse dictionary offers the user to enter the phrase “moving down a steep slope by holding on to a rope” as input, and expect to receive the word “abseil” as output. The reverse dictionary finds out the efficient result by assigning each thread for each word phrase.

The reverse dictionary addresses the widespread problem of “word is on the tip of my tongue, but I can’t able to get that word”. A particular category of people afflicted heavily by this problem are writers, students, teachers, scientists, professional writers, advertisement professionals, marketing professionals, etc. Many people with certain level of education, the problem is not lacking knowledge of words, but unable to recall the appropriate words. The most effective technique applied to converse mapping is latent semantic indexing (LSI) and principal component analysis (PCA).

Latent Semantic Indexing is a dimensionality reduction technique which reduces the descriptive length of the document statistical structure. Principal Component Analysis is a statistical procedure that converts the correlated variables into a set of values of linearly uncorrelated variables.

Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Indexing (PLSI) are used for building efficient schemes in reverse dictionary. LDA finds the short description from a large collection of document while preserving the essential statistical relationship. The approach of LDA reduces the descriptive length of document that in turn provides modularity and extensibility. PLSI defines a generative model of the data and minimizes word perplexity. The standard techniques from statistics can be applied for combining the models and controlling the complexities. The reverse mapping set maps for all terms appearing in the sense phrase and determine the most generic form by reducing various inflexions of the same word to a common form.

The converse mapping technique uses stepwise refinement approach for extracting the best matched result for an input phrase. The approach on information retrieval finds the core terms from the sense phrase and searches for the candidate words whose definitions containing the similar core terms. The synonyms, hyponyms and hypernyms of the terms increase the probability of obtaining sufficient number of outputs. The results are sorted by comparing the input with every definition in the dictionary database. The similarity measure between the sentences determines the importance of terms that contribute more to the meaning of a phrase. In modern information retrieval, the converse mapping technique can provide significant improvements in performance and extracts efficient result without impacting available solution quality.

II. RELATED ARTICLES

In a recent study, Blei et al [1] proposed a generative model for collection of finite data. Latent Dirichlet Allocation is a three level hierarchical Bayesian model, that takes the relevant data from a large collection of document while preserving the originality and provides explicit representation of a document. The efficient inference techniques based on variational method estimates the Bayes parameter. Carlberger et al [2] evaluated a linguistic normalization technique called stemming in which the variant forms of a word are reduced to a common form.

Manuscript published on 28 February 2014.
* Correspondence Author (s)
M. Ramya, Department of Computer Science and Engineering, Sathyabama University, Chennai, India.
E. Thenmozhi, Department of Computer Science and Engineering, Sathyabama University, Chennai, India.

For example, stemming transforms the variant forms such as 'connection, connections, connective, connected and connecting' to a common form called 'connect'. Stemming improves both precision and recall in information retrieval. Precision is the number of relevant retrieved document to the retrieved document. Recall is the number of relevant retrieved document to the relevant document.

Dao et al [3] measured the similarity between the meaning of two sentences by using the semantic relation, lexical relation and parts of speech. Semantic relations include hyponym, hypernym and troponym. Lexical relations include synonym and antonym. The semantic similarity between the two sentences can be computed by using the techniques such as tokenization, tagging, stemming and word sense disambiguation approach. Ponte et al [4] designed a language modeling approach that estimates models for each document individually. The approach to modeling is non-parametric and integrates document indexing and document retrieval into a single model. The probability of generating the query is estimated and the documents are ranked according to the probabilities. The language models can provide accurate representation of the data which is simple and explanatory.

Earley et al [5] generated an efficient context-free parsing algorithm. The efficiency of the algorithm is used to achieve the time and space bound in the natural language processing system. The context-free parsing algorithm appears to be superior to the top-down and bottom-up algorithms. The semantic routines can be associated with productions without reflecting the original structure of the grammar. Callan et al [6] proposed the passage-level evidence in document retrieval that helps to understand how the passages should be defined, ranked and retrieved using the approaches such as paragraphs, bounded paragraphs and fixed-length window. The advantage of passage level evidence while retrieving the information includes precision, efficiency and simplicity.

Hofmann et al [7] developed an approach to automated document indexing and information retrieval called latent semantic indexing that attempts to overcome the problems by mapping documents based on terms. The specific form of mapping is determined by document collection and based on Singular Value Decomposition of the document matrix. Probabilistic Latent Semantic Indexing defines a statistical model that replaces the original vector space representation of the documents by a low dimensional latent space representation. Lafferty et al [8] designed the information retrieval framework that combines document models and query models using a probability ranking function based on Bayesian decision theory. A language model for each document and each query is estimated, and the retrieval problem is cast in terms of risk minimization.

Wu et al [9] evaluated the semantic representation that defines each verb by a set of concepts in different conceptual domain. Based on the conceptual representation, a similarity measure allows to achieve correct lexical choice even when there is no exact lexical match from the source language to the target language. The verb semantic representation has a great impact on the quality of lexical selection. Zwick et al [10] proposed the similarity among objects as a linear combination of their common and distinct features. The generalization depend on the definitions of cardinality and difference in fuzzy set theory. The efficiency of the linguistic approximation process represent each fuzzy set by a limited number of features so that the distance computation is simplified.

III. EXISTING SYSTEM

The reverse dictionary is unlikely to match the exact result for the user entered keyword from the server database. The converse mapping technique returns a set of possible matches from which a user may select the choice of terms. The mapping is complex and the user is unlikely to enter a definition that exactly matches with the dictionary definition. The efficiency to respond to a request is less and the end user becomes impatient, if a website takes longer time. The reverse dictionary needs to be usable online. The key requirement in the performance of reverse dictionary is to allow online interaction with users. Current semantic similarity measurement schemes are highly intensive, making online responsiveness and scaling of converse mapping difficult.

The CSP is a well-known hard problem that makes the existing results infeasible. The concept similarity problem significantly represents the same word with different characteristics that has been addressed in a variety of fields such as psychology, linguistics, and computer science. For example, the word "interest" specifies different meanings with respect to subject and banking. Semantic similarity is more for single word concepts and much less for multiple word concepts. The concept similarity problem leads to difficulty in analyzing and estimating the concepts. In reverse dictionary, the documents considered for similarity measurement are very short dictionary definitions which contain very little contextual information. The lack of contextual information in the reverse dictionary adds to the difficulty of addressing the problem space.

IV. PROPOSED METHODOLOGY

The reverse dictionary system is based on the notion that returns the set of possible matches in the form of word phrase as output for a given sense phrase. For example, the reverse dictionary takes the description of concept "fact that is not pleasant and difficult to accept" as input and returns set of possible matches such as "swingeing, unpalatable, abrupt" as output. The user faces difficulty in finding the exact match from the list of possible matches. The complexity in reverse dictionary can be solved by using the comparison process. The converse mapping technique in reverse dictionary matches the exact word as output that resembles the description of desired concept given in the input phrase. For example, the reverse dictionary takes the description of concept "fact that is not pleasant and difficult to accept" as input and returns the exact candidate word "unpalatable" as output. Entering the sense phrase that exactly matches with the dictionary definition is complex. Such potential matches can be generated by comparing the user input phrase with every definition in the dictionary. The approach on comparing the keywords faces two major problems. First, the input given by the user should exactly match with the definition in dictionary. (i.e) The user input phrase must contain the same keywords as that of dictionary definition. Second, the user faces difficulty in comparing the input with all the words in the dictionary, if a dictionary contains more than one lakh words. The major problems in the comparison process can be solved by reducing the set of definitions and query expansion technique.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The method of increasing the number of terms in the query in order to find the relevant output is called query expansion.

The efficiency of the reverse dictionary can be improved by reducing the number of definitions in the comparison process. Indexing technique limits the large amount of information and takes only the relevant data from the database. The reverse mapping index maps the keywords in the input phrase with the definitions in dictionary containing the same keyword. Stemming reduces various forms of a word to a common form. Similarity of concept with different grammatical meaning can be avoided by the stemming technique. Reverse mapping index and stemming are used to limit the number of definitions during the process of comparison. The query expansion technique increases the potential of reverse dictionary to identify the relevant definition when there is no exact match from source definition given by user to the target definition in dictionary.

A. *System Architecture*

The design and implementation of reverse dictionary application is represented in fig.1. The reverse dictionary application takes the sense phrase as input and partition the sense phrase into separate words. The stop words are removed to increase the overall scalability of the system. The negation method combines the negation word set with the candidate word in order to convert two different forms of a word to a common form. The core terms of the input phrase is obtained by implementing the stemming technique. The most generic forms of the words are found by allowing the terms to undergo the process of stemming. Thread is allocated for each core terms and every thread is combined to form the thread pool. The thread pool allows parallel retrieval of data from the databases that includes reverse mapping set database, similarity database, synonym database, hypernym database, hyponym database and definition database.

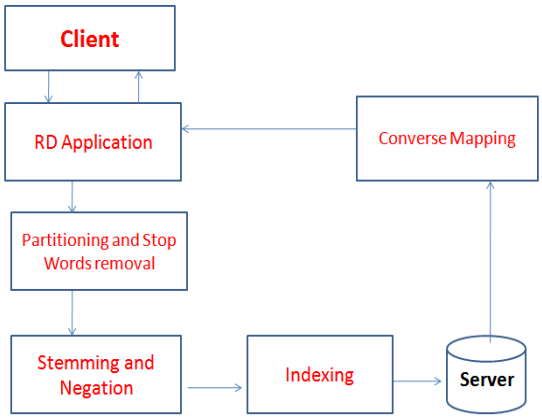

Figure1: Proposed Block Diagram

Cache is a temporary memory that maintains a database which consists of frequently accessed data. Indexing is used to map the core terms of the input phrase with the available databases and searches for the relevant data. When the user gives the input phrase in the reverse dictionary application, the thread pool checks the cache whether the appropriate data is available. If the data is available, the cache returns the required data the thread. If the required data is not available, the cache returns null set to the thread. The thread pool then contacts the databases and receives the relevant data from the databases by applying the indexing technique. The converse mapping technique maps from sense phrase to appropriate word phrase. The results of converse mapping are sorted in decreasing order by comparing the input phrase with the dictionary definition. The reverse dictionary application returns the top word phrase in the ranking process as the exactly matched output.

B. *Stepwise Refinement Approach*

The stepwise refinement approach uses converse mapping technique to extract the efficient result for the user entered input phrase without affecting the original quality. The approach on refining the best matched word phrase consists of following steps: 1)Partitioning the sense phrase 2)Stop words removal 3)Replacing negation words 4)Stemming 5)Indexing 6)Ranking.

1)Partitioning the sense phrase: Word phrase refers to a single term that denotes specific meaning. More number of terms are combined to form a sense phrase. The user input phrase for the reverse dictionary consists of sequence of terms that are usually called as sense phrase. The reverse dictionary application partition the sense phrase given as input into separate word phrases. For example, the reverse dictionary takes the input as “exploring natural caves” and partition the sense phrase into sequence of terms such as “exploring, natural, caves”. The partitioning method increases the overall scalability of the system by allocating separate thread for the word phrases. The thread specifies the concurrent execution of partitioned words. The thread pool is created by combining the entire threads and allows to access the information from the databases.

2)Stop words removal: The words which does not make any change in the output while removing is called stop words. The stop words are removed from the query after partitioning the user entered input phrase. The stop words consists of two different levels such as level 1 stop words and level 2 stop words. The first level of stop words are not useful in finding the output. Such stop words that cannot be used for the purpose of indexing are always removed from the input query. The second level of stop words may or may not be useful in finding the output. Depending on the usefulness in indexing, the words proceed to the next step in the refinement approach. The stop words removal approach is used to achieve the time bound and space bound in reverse dictionary during the process of execution.

3)Replacing negation words: The replacement technique combines two different terms, one of which containing the negated set to form a meaningful word phrase. For example, the negation word “not” and word phrase “expected” are combined to form a meaningful word “unexpected”. The negation word set phrases are replaced by antonyms which contains the same conceptual meaning. The synonyms of the newly derived word are also taken under consideration for the purpose of indexing. For example, the word “unexpected” containing the synonym “unforeseen” are also used for indexing. The approach on negation generates sufficient number of terms for extracting the relevant matches. The query expansion technique is used to improve the result set of the reverse dictionary.

4)Stemming: Stemming is a technique to obtain the most generic form of a word. (i.e) Different inflections and derivations of the same word can be transformed to a common form.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

For example, different inflections of the same word "relative, related, relation, relating, relativity and relations" can be transformed to a common form called "relate". Stemming is also used to retrieve the information by removing the prefix and suffix of a word. The information retrieval system uses a traditional technique such as stemming and normalization for both the query and text. Stemming can give better precision in information retrieval for short queries on short documents. Precision depends on the query length and document length. The motivation for using stemming instead of lemmatization is the question of cost in terms of time and effort.

5)Indexing: The reverse dictionary applies the method of indexing in the refinement process. Indexing is an approach that takes the partitioned, stop words removed, negated and stemmed candidate words as input and checks for similar words in the dictionary. The candidate words must occur simultaneously in each of the dictionary definitions obtained during the process of indexing. If the dictionary does not generate sufficient number of results, several relations of the extracted candidate words are taken under consideration. The relationship among words includes synonym, antonym, hypernym, hyponym and troponym. Indexing technique searches for the related words in the dictionary definitions in order to obtain the efficient result. The process of finding relevant definition from the dictionary database continues until the indexing technique generates sufficient number of results.

6)Ranking: The ranking phrase compares the input given by the user with the dictionary definitions obtained during indexing and rank the generated results based on the semantic similarity function. Measures of such similarity can be found using techniques such as term similarity, term importance and weighted similarity measure. Term similarity is used to find the similarity measure between the terms. Term importance describes the importance of a term in a particular phrase. Weighted similarity measure determines the weight for every pair of terms. The results generated while indexing are sorted in decreasing order using the ranking function and the top result is considered as the exact output. The stepwise refinement approach takes the sense phrase as input and returns the efficient word phrase that satisfies the description of concept.

V. CONCLUSION

This paper presents an efficient converse mapping from sense phrase to word phrase. We propose set of techniques to find the exact result for the entered sense phrase, when the user is unable to recall the appropriate word on demand. In this paper, the exact candidate word can be found by partitioning and comparing the input phrase with the dictionary database. The query expansion technique increases the efficiency and scalability of reverse dictionary. The converse mapping technique in modern information retrieval provide significant improvements in the performance scale and extracts the best matched result without sacrificing the quality of the solution.

REFERENCES

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation", *J.Machine Learning Research*, vol. 3, pp. 993-1022, Mar. 2003.

[2] J. Carlberger, H. Dalianis, M. Hasssel, and O. Knutsson, "Improving Precision in Information Retrieval for Swedish Using Stemming", Technical Report IPLab-194, TRITA-NA-P0116, Aug. 2001.

[3] T. Dao and T. Simpson, "Measuring Similarity between Sentences",http://opensvn.csie.org/WordNetDotNet/trunk/Projects/Thanh/Paper/WordNetNet_Semantic_Similarity.pdf, 2009.

[4] J. Ponte and W. Croft, "A Language Modeling Approach to Information Retrieval", *Proc. 21st Ann. Intl'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pp.275-281, 1998.

[5] J. Earley, "An Efficient Context-Free Parsing Algorithm", *Comm.ACM*, vol. 13, no. 2, pp. 94-104, 1970.

[6] James P. Callan, "Passage- Level Evidence in Information Retrieval", *Proc. 17th Ann. Intl'l ACM SIGIR Conf. Research and Development in Information Retrieval*, July 1994.

[7] T. Hofmann, "Probablistic Latent Semantic Indexing", *SIGIR'99: Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 50-57, 1999.

[8] J. Lafferty and C. Zhai, "Document Language Models, Query Models, and Risk Minimization for Information Retrieval", *Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 111-119, 2001.

[9] Z. Wu and M. Palmer, "Verbs Semantics and Lexical Selection", *Proc. 32nd Ann. Meeting Assoc. for Computational Linguistics*, pp. 133-138, 1994.

[10] R. Zwick, E. Carlstein, and D.Budescu, "Measures of Similarity Among Fuzzy Concepts: A Comparative Analysis", *Int'l J. Approximate Reasoning*, vol. 1, no. 2, pp. 221-242, 1987.