An Information Retrieval Method for Digitally Communicated Datasets to Address High Recall Requirements

Venkata Krishna Kota and Umamaheswaran S

Central Research Laboratory, Bharat Electronics Limited,
Bangalore, India

Abstract- Emails, SMS, Chat transcripts and Social Network Messages play a vital role in present day's communication. These digitally communicated data sets are different from well typed English documents like books, news papers …Etc in 2 ways. One is the use of acronyms instead of typing full form of the term and another one is the existence of misspelled words. Because of these reasons, information retrieval methods used for analyzing well typed English documents may not be well suitable to analyze them. In this paper we propose an information retrieval method to address the high recall requirements of retrieving information from digitally communicated data sets.

Index Terms- Information Retrieval; Search Engine; Index Term Expansion; Semantic Search

I. Introduction

Information retrieval plays a vital role in this information age. Information retrieval is the process of retrieving information or documents that contain information or metadata about documents which is relevant to the given query from huge data collections [9]. Indexing is a key step in information retrieval. Indexing refers to the method of processing the original data into a highly efficient cross-reference lookup in order to facilitate rapid searching. While indexing, content from documents will be acquired, the acquired content will be tokenized into index terms and these terms will be indexed.

Emails, SMS, chat transcripts, through social networking sites etc are popular ways of communications among people today. There is huge amount of such digitally communicated data exist. There is a thirsty need for tools to search such data. These digitally communicated data sets are different from well typed documents like books, research articles, news etc. Documents like books, news, articles etc will have proper sentences. They are grammatically correct. The words in sentences are spelled properly. But it is not the case with digitally communicated data. For the sake of convenience, people normally use acronyms instead of typing full words while they are digitally communicating with others. For example they use '*tc*' instead of '*take care*'. They may not focus on the spelling of the words because in most of the times human can understand the semantics behind a misspelled word. For example '*recieved*' can be identified by human as '*received*' easily.

In the case of well typed documents, the index terms extracted while indexing are mostly proper terms i.e they are properly spelled words. But in the case of digitally communicated data sets, the index terms can be chat acronyms, misspelled words. Assume that these acronym and misspelled index terms are indexed as they are. When user submits a query, even if a document is relevant to the given query it may not be retrieved. For example an instant message contains the term '*tc*' and it is indexed. The intension of using '*tc*' is to convey '*take care*'. When user submits a query '*take care*', that instant message will not be retrieved because '*take care*' is not an index term for that instant message. Thus even if a document is relevant, it may not be retrieved. Dealing with acronyms and misspelled words are major challenges for indexing digitally communicated data sets. In this paper we present a method for indexing digitally communicated data sets. The proposed method applies index term expansion for acronyms and misspelled words. If an index term is an acronym, then that acronym and its expansion terms will be indexed. If an index term is a misspelled term, then that misspelled term and possible correctly spelled terms for that misspelled terms will be indexed.

II. Related Work

In this digital age, emails, SMS, Chat transcripts, Social network messages and other kinds of digital communications play a vital role. The volume of such data is huge in nature. They possess potential information. This information is useful for many applications. In this paper a method is proposed to retrieve information from digitally communicated data sets.

Information Retrieval is the process of retrieving information or documents that contain information or metadata about documents relevant to the given query. Keyword based search tools plays a vital role in many applications. To preserve the semantics of keyword based search systems there are 2 main approaches. They are "Query term expansion" and "Index term expansion". These are possible with the help of lexical reference systems [9]. With these approaches recall of the search will be improved. But it may affect precision of the search. Special ranking methods can be applied to handle precision degradation. Term expansion is favorite choice in applications like "Digital Forensic Search" which have high recall requirements.

WordNet is an online lexical reference system [8]. It contains English lexicons and semantic relationships among those lexicons. It is widely used by many researchers in their experiments. WordNet divides lexicon into 5 categories: nouns, verbs, adjectives, adverbs and functional words. It covers the vast majority of nouns, verbs, adjectives and adverbs from English.

Words in WordNet are organized in synonyms sets called synsets. Each synset represents a concept [4, 2].

Query Term Expansion is the process of expanding the Query terms with their semantic equivalents. Query Term Expansion is done in many applications to preserve the semantics of the retrieval functionality. In Query Term Expansion, the user query will be analyzed and expanded with semantic equivalents of query keywords [1]. With Query term expansion, index space can be optimized with the cost of extra query processing time. In [4], author used WordNet ontology for query expansion to achieve high recall requirements. Query terms have been expanded with their synonyms, and other related terms in the same domain. In [3], Query Expansion technique is proposed which considers phrases as its expansion unit. In this paper also query term expansion is used with the help of WordNet ontology**.**

Index Term Expansion is the process of expanding the index terms with their semantic equivalents.Index term expansion will do term expansion at index time. Through this process we can optimize search time with the cost of additional index space. Through index term expansion, along with the actual terms in that document, their semantic equivalents also will be indexed. Thus semantics of the search can be preserved. In general Index Term Expansion will be done to expand the index terms with their synonyms. In [5] authors used a method based on the expansion of index terms, which exploits WordNet synonyms and holonyms. It is used in order to find implicit geographic information from text, particularly in the cases in which the indication of the containing geographical entity is omitted.

These approaches suits well for well typed documents like books, news papers, web pages .etc. But they may not work well with email, SMS .etc because they don't bother about usage of acronyms / short form words and existence of misspelled words while indexing.

In this paper a method is proposed which address these difficulties.

III. METHODOLOGY

Existing tools do term expansion to expand the index terms with their synonyms and other related terms. If the indexing term which needs to be expanded itself is not a proper English word, then obviously they cannot get its semantic equivalent terms because the underlying lexical dictionary does not possess that term. But if one observes the nature of emails, chat transcripts, SMS .etc, most of the terms are not proper English terms. In this paper a method is proposed to index such documents.

Use of acronyms is a common approach while communicating through emails, SMS, chat transcripts, Social network messages …etc. For example people use '*gn*' instead of '*good night*' for the sake of easiness. Human can easily understand the semantics behind these acronyms. Information Retrieval tools developed to index well typed documents may not get the intension of using those acronyms. So they index those acronyms as they are and fail to capture the semantics of the document.

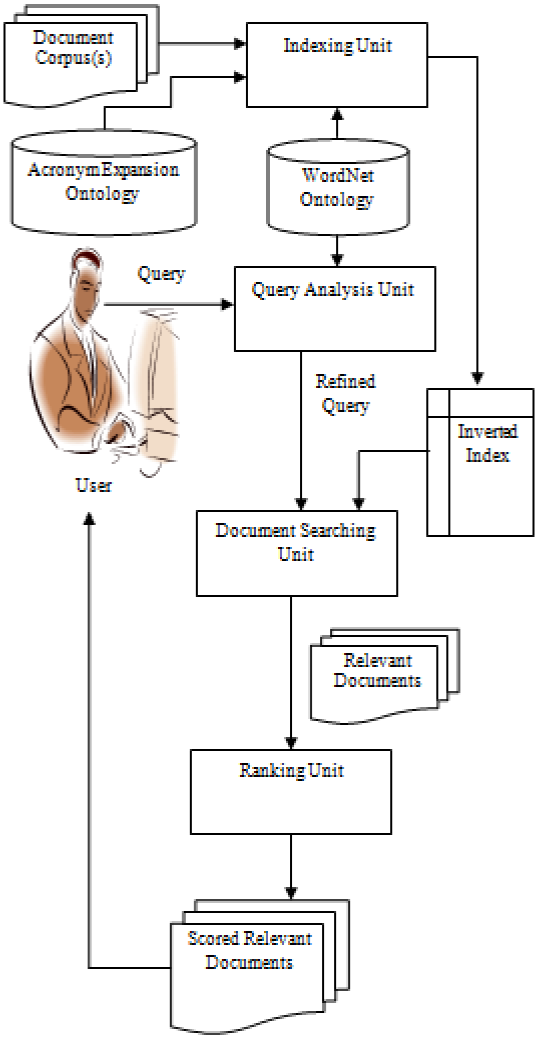

Figure1. Block diagram of Proposed System.

Unlike well typed English documents like books, news papers …Etc, documents like emails, chat transcripts, social network messages contain many misspelled words in it. Main intension of people is to convey the message to others with ease. For the sake of ease they use acronyms. They will not bother whether the terms are properly spelled or not. If the other person can get the intension behind the misspelled term, there need not to bother. Even if there exist, slight spelling mistakes, the person who is reading it can identify it as a misspelled word and can guess the actual word. If the indexing process cannot capture the semantics behind the misspelled terms, it will index them as it is and it cannot be retrieved even if it is relevant to the user's query. So the indexing process should be capable of identify misspelled words and should perform index term expansion for misspelled words.

The block diagram of the proposed system is shown in Fig 1.

Each document in the document corpus is analyzed by the Indexing unit. The concept of processing the original data into a highly efficient cross reference lookup in order to facilitate rapid searching is called indexing [6, 7]. Indexing contains multiple phases like acquiring content from documents, parsing, tokenizing, expanding index terms and adding the terms to the inverted index. Indexing unit will acquire content from the document, parses it, analyzes it, tokenizes it and adds it to the inverted index.

Indexing algorithm is given below.

```
ALGORITHM indexCorpus(coupus, WorNet, AEO)
//Description: indexes the given corpus with the help of WordNet and AEO returns Inverted_Index
//Input
// corpus is the document corpus that we need to analyze and index
//WordNet is a lexical reference system
//AEO is proposed Acronym Expansion Ontology
//Output
//Output is the Inverted_Index
{
Create an empty Inverted_Index
For each document in the corpus, DO
{
    Analyze the document and generate the Token_Stream
    FOR each token in the Token_Stream, DO
    {
        Lowercase the token
        IF that token exist in AEO, THEN
        // probably it represents an acronym
        {
            Get the acronym expansion terms for that token from AEO
            Expand the index terms with acronym expansion terms for that token
        }
        Else
        {
            IF that token does not exist in WordNet, THEN
            //probably it may be a misspelled term
            {
                Get set of terms from WordNet such that having lesser String_Distance (less than the user configured value) to that token and say they set as S1
                Get set of terms from WordNet such that they have the soundex value equivalent to the soundex value of that token and say that set as S2
                Expand the index terms with terms in (S1 U S2)
            }
        }
  }
Add all the index terms of that document  to the inverted index
}
Return Inverted_Index
}
```

For expanding index terms which are acronyms we propose Acronym Expansion Ontology. The Acronym Expansion Ontology contains acronyms and expansions for them. Every index term will be checked in this ontology. If the term exists in the ontology then it will be identified as an acronym. Corresponding acronym expansions for that acronym will be extracted from ontology. The acronym term and extracted acronym expansion terms will be indexed.

WordNet Ontology contains most of the English terms and semantic relations among them. Misspelled terms will be identified using WordNet ontology. If an index term does not exist in WordNet Ontology, then there is more chance for it to be a misspelled word and it is identified as misspelled word. For each misspelled term, nearest correctly spelled terms will be identified. Each of this nearest correct term is a term from the WordNet such that the string distance between it and the misspelled term is small or it is phonotonically similar (having the same soundex) with the misspelled term. The misspelled term and its nearest correct terms will be indexed. Once every document in the document corpus is analyzed and indexed, the system is ready for a search operation.

When user queries a system, the query will be analyzed, expanded and refined by the Query Analysis Unit. Query Analysis Unit refines the query by expanding the query terms with their semantic equivalents with the help of WordNet Ontology. This refined query is given to the Document Search Unit which retrieves the documents that are relevant to the refined query. This set of relevant documents will be fed to the Ranking Unit. Ranking unit calculates relevance scores for each retrieved document based on their relevance to the given query and ranks the retrieved documents based on their relevance scores. These Scored Relevant Documents will be displayed to the user.

IV. CONCLUSION

An indexing method is proposed for indexing digitally communicated data sets like emails, SMS, instant messages and social network messages .etc. Usage of acronyms and spelling mistakes are most common in such data sets. The proposed method applies index term expansion while indexing acronyms using proposed Acronym Expansion Ontology. The proposed method applies index term expansion while indexing misspelled words using WordNet Ontology to identify misspelled words. It uses string distance and soundex techniques to identify the probable spell corrections. With this indexing method high recall requirements of retrieving information from digitally communicated data sets can be achieved.

REFERENCES

[1] Hazra Imran and Aditi Sharan, "Thesaurus and Query Expansion", International Journal of Computer Science & Information Technology, 2009, vol 1, No 2, pp. 89-97.

[2] Christiane Fellbaum, "WordNet" , Theory and Applications of Ontology, 2010, pp. 231-243, doi:10.1007/978-90-481-8847-5_10

[3] Yongli Lium, Chao Li, Pin Zhang and Zhang Xiong, "A Query Expansion Algorithm based on Phrases Semantic Similarity", 2008, International Symposium on Information Processing, pp. 31-35, doi: 10.1109/ISIP.2008.57

[4] Report, Australian Phan Thien Son, "Ontology-Driven Text Mining for Digital Forensics", COMP6703 Project National University, 2007.

[5] Davide Buscaldi, Paolo Rosso and Emilio Sanchis, "WordNet-based Index Terms Expansion for Geographical Information Retrieval", Evaluation of Multilingual and Multi-modal Information Retrieval, 2007, pp. 954-957, doi:10.1007/978-3-540-74999-8_122

[6] Erik Hatcher, Otis Gospodnetic and Michael McCandless, “Lucene in Action, Second Edition”, Manning Publications, 2009.

[7] Lucene search library, available at http://lucene.apache.org/nutch

[8] WordNet, available at http://wordnet.princeton.edu

[9] Christopher D Manning, Prabhakar Raghavan and Hinrich Schutze, “Introduction to Information Retrieval”, Cambridge University Press, 2008

AUTHORS

First Author – Venkata Krishna Kota, Central Research Laboratory, Bharat Electronics Limited, Bangalore, India
Email id - venkatakrishnav@bel.co.in

Second Author – Umamaheswaran S, Central Research Laboratory, Bharat Electronics Limited, Bangalore, India