Query processing and data extraction using word embedding technique and VSM

Purudewa Pawar
Master of Technology
IIITD
New Delhi, India
purudewa20053@iiitd.ac.in

Kapil Bhargawa
Master of Technology
IIITD
New Delhi, India
kapil20088@iiitd.ac.in

***Abstract*—In today's growing world, retrieving the useful information form the large pool has become crucial task to avoid access delay and filtering unwanted information.The information retrieval systems are responsible for extracting the useful information and shows the relevant documents against the input query to the user. This work includes implementation of various models in information retrieval system in order to improve performance and produce the most relevant outputs. BM25, VSM etc are Incorporated in the system for execution.**

***Index Terms*—Information retrieval, Similarity, BM25, VSM, Word embeddings**

I. Problem Statement

The problem defines producing the search results for the user query which may be related to specific domain or document. Problem arises when user would like to retrieve information for a particular product or if user wants to extract information from a restricted content of data. This scenario can be related to when user would like to retrieve information from a particular document or suppose there is an e-commerce website and user would like to search for a product which pertains to their need. In such cases, a general solution would be to simply google. But it does not sound much great to search for a solution's problem which is restricted to a particular domain. Google will be giving generalized result.

Fig. 1. Finding specific info from a large content.

A. Reason to solve problem

Building a system to extracting or obtaining search results that user or company tends to obtain is more efficient as it avoid the involvement of other party applications. It doesn't seems fit to involve third party application in order to address problem pertaining to our company. Moreover, user's private data cannot be shared directly with the third party in order to address the problem.

B. Task

In these scenarios, we as a developer need to create our own model which would be able to address the problems pertaining to a particular domain. This is the point where information retrieval comes into play. We need to retrieve information in an efficient manner which would suit the users best interest. Representation of user query in best possible manner and retrieving the most efficient matched results is the purpose of IR systems.

C. Dataset used

The corpus of documents used in this work contains 450 documents, 100 testing queries and top 10 relevant documents ids corresponding to the queries.

II. Literature Review

During our analysis on existing models, we got to know about various existing techniques used in retrieving the relevant documents with respect to the query. Existing models such as VSM (VECTOR SPACE MODEL), BM25 MODEL, ELASTIC BM25 MODEL. There were other mathematical model which simply used some similarity metric in order to find the relevant document with respect to the query. A fixed evaluation metric was not known for the above models since the performance of model may vary depending on the type of dataset present.

A. Combining Word2Vec with Revised Vector Space Model for Better Code Retrieval

The paper aims at proposing better techniques for API code examples retrievals. Traditionally there were techniques that were used that are information retrieval techniques with text matching that computes the textual similarity of the query with API codes, explemer that takes query and use IR and program analysis to find relevant code examples etc. But these API code retrieval methods suffer from a problem that is a lexical mismatch problem in which a developer of code may use different names that may be different from english language queries. Word2Vec addresses the lexical mismatch problem in code search/retrieval. The Word2Vec is executed on English term and API code elements including classes and methods. The Word2Vec wil have shared vector space with

English terms and API code elements. Semantic similarity then calculated between the query and API code is measured based on the distances of the vectors of terms in query and code elements in API code example.

B. Toward Incorporation of Relevant Documents in word2vec

The term relatedness plays a crucial role in information retrieval that includes query and document matching. The embedding models in general define the term relatedness by exploiting the terms' co-occurrences in short-window contexts. The approach is to find related terms to query using local information i.e. a set of top retrieved documents. This paper focuses on incorporating the local information of the query in the word embeddings. One main challenge in this direction is the dense vectors of word embeddings and their estimation of term-to-term relatedness remain difficult to interpret and hard to analyze.The solution/method proposed in this paper includes explicitwordrepresentationsproposevectors whose dimensions are easily interpretable. One fundamental issue in using word embedding (or most other methods of term similarity) is rooted in the assumption of independence of query terms i.e. the similar terms to a query term are independent of the other query terms. This issue is resolved by training separate embedding models on local information of the query i.e. a set of top retrieved documents.They show that the locally-trained word embeddings outperform the global embedding model as the similar terms are relevant to the whole query.

C. Toward Word Embedding for Personalized Information Retrieval Word Embedding for Personalized Information Retrieval

In the information retrieval methods using Word2Vec, word embeddings are learned on a general corpus that has a huge number of documents. This paper focuses on personalizing the word embeddings learning, by achieving the learning on the user's profile. The word embeddings are then in the same context than the user interests. This method of word embeddings exploits the local information and doesn't use global embeddings. Similarly to Latent Semantic Analysis (LSA), Word Embedding maps the words to low-dimensional (w.r.t. vocabulary size) vectors of real numbers. For example, two vectors t0 and t1, corresponding to the words t0 and t1, are close in a N-dimensional space if they have similar contexts and vice-versa, i.e. if the contexts in turn have similar words. The idea of the paper: Select words that occur in the same context as the terms of the query. We compare then Word Embedding learned on the whole collection of Social Book Search, called the Non Personalized Query Expansion, versus Word Embedding learned on the user's profiles, called the Personalized Query Expansion.

D. Application of Vector Space Model to Query Ranking and Information Retrieval

This paper addresses the issue of database volatility in which documents can be added, modified and deleted in large databases thus making it difficult for information retrieval systems. This issue also restricts the usage of data structure and algorithms for application. This paper focuses on the Vector Space Model (VSM) technique of information retrieval. First, compute the similarity scores using the weighted average of each item. The cosine measure is then used to compute the similarity measure and to determine the angle between document's vector and the query vector since VSMs are based on geometry whereby each term has its own dimension in a multidimensional space, queries and documents are points or vectors in this space. The cosine measure is often used. Then it is easier to retrieve data or information based on their similarity measures and produces a better and more efficient technique or model for information retrieval.

E. Evaluating vector-space models of analogy

The aim is to compute the similarity between the objects/documents or between the query and database terms.Vectorspace representations provide geometric tools for reasoning about the similarity of a set of objects and their relationships.Recognizing that two situations have similar patterns of relationships, even though they may be superficially dissimilar, is essential for intelligence. This ability allows a reasoner to transfer knowledge from familiar situations to unfamiliar but analogous situations, enabling analogy to become a powerful teaching tool . The vector space used in the paper is capable of capturing the verbal analogies. The paper evaluates the parallelogram model of analogy as applied to modern word embeddings, providing a detailed analysis of the extent to which this approach captures human relational similarity judgments in a large benchmark dataset. However, the questions of where these relations come from and how to determine that the relationship between one pair of entities is the same as that between another pair is still not resolved. In this paper approach, entities are represented as points in a Euclidean space and relations between entities are represented as their difference vectors. Even though two pairs of points may be far apart in the space (i.e., they are featurally dissimilar), they are considered relationally similar as long as their difference vectors are similar. This paper evaluates the parallelogram model of analogy as applied to modern vector-space representations of words.

F. BM25 : A ranking function for information retrieval systems

BM25 is a ranking function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters. It is used by search engines to rank matching documents according to their relevance to a given search query and is often referred to as "Okapi BM25," since the Okapi information retrieval system was the first system implementing this function.

G. Application of Output Embedding on Word2Vec

The word vector of distributed representation that embeds the semantic relationship of words into a vector using

Word2Vec has attracted attention in recent years. Furthermore, this word vector has become widely to be used in the field of Natural Language Processing such as parsing and document classification, and its effectiveness has been reported. Generally, it is the input embedding on Word2Vec that is used as the elements of a word vector, and the output embedding generated at the same time is not used. On the other hand, the authors focus on the usefulness of the paired output embedding. In this paper, we propose word vectors using the input and output embeddings together. In addition, we experimentally investigate the performance of the proposed word vectors and carry out document classification experiments using the proposed word vectors. The result shows that the classification performance was improved by the proposed word vectors.

H. A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction

With the steady accumulation of unstructured data, the domain of natural language processing (NLP) is gaining widespread attraction amongst researchers and practitioners in order to quickly and easily extracts prediction-like insights in a simplified and streamlined fashion. The subject of text mining and analytics is going through a variety of delectable advancements. There are a number of articulations on the subjects of NLP and machine learning (ML). Very recently, the model of the bag of words has become so popular in order to produce accurate predictions out of unstructured text data. In this paper, we have explained an easy-to-use framework for accelerated usage of the BoW model towards pioneering text mining and processing. We have demonstrated a simple example by leveraging this framework in order to showcase the utility of this generic framework that can be easily replicated across in manyother associated scenarios.

III. Baseline Model Evaluations

A. Information Retrieval using Vector Space Model technique.

Our first baseline model is VSM technique. In Vector Space Model, each document or query is a N-dimensional vector where N is the number of distinct terms over all the documents and queries. Value present at i-th index in vector represent the TF-IDF score for the i-th term in vector. We have used Term-Frequency (tf) and Inverse-Document Frequency(idf) to evaluate the score. After building up the vector space model, now we can fetch the relevant documents to our query using the cosine similarity measure. Now we have sorted the retrieved documents in decreasing order of similarity. We have used Mean Average Precision(MAP) value as our metric for evaluating the IR model. We are having the ground truth values for our query and accordingly average precision value can be evaluated for each query. After getting the average precision value for each query, we have evaluated the Mean Average Precision(MAP) value for the whole IR model.MAP value reported for VSM model is 0.67.

B. Information Retrieval using BM25 technique.

Our second baseline model is BM25 technique. BM25 model tries to overcome the TF-IDF model by mapping the relevance as a probability problem . A probability score for a particular document indicates the relevance score of document for a given query. For a given particular query, using our BM25 model, we can evaluate the relevance score for documents. Now we have sorted the documents in decreasing order of their relevance score. As we have used the Mean Average Precision(MAP) metric for VSM model, similarly it is used to evaluate BM25 technique also. MAP value reported for BM25 model is 0.59.

TABLE I
BASELINE MODEL RESULTS

Baseline model	***MAP value***
Vector space model	0.67
BM25	0.59

IV. Finally Proposed Model

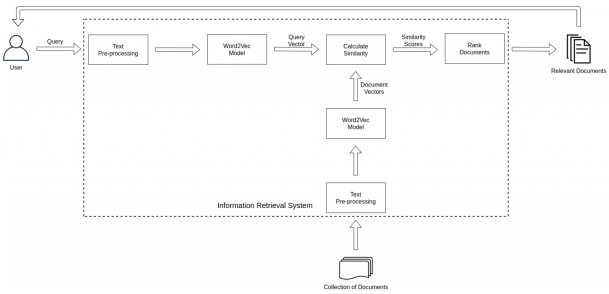

Fig. 2. Proposed methodology.

Proposed model methodology

In baseline models, we had created the unique vocabulary and then evaluated the scores for the documents.In our proposed method, We have implemented word embedding techniques such as BOW(BagOfWords) and Word2vec in order to improve the MAP value.We have implemented these two word embedding technique in VSM model and accordingly the MAP value is reported. "WORD2VEC" model converts words to vectors.Promising feature of WORD2VEC model is it's ability tocapture context of data and represent it using vectors. Thus,it is able to maintain the semantic and syntactic association between words. We arrived at the conclusion that the Word2Vec word embedding technique gave good results as compared to BagOfWords technique. MAP value reported using Word2Vec embedding technique is 0.74 which is improvised over baseline models.

V. Conclusion

It has been observed that the finally proposed model has performed better than the given baseline model. The MAP

TABLE II
FINALLY PROPOSED MODELS RESULT

Model name	***MAP value***
BOW model	0.48
Word2Vec model	0.74

measure is more for our proposed method in comparison to the baseline models. The word2vec word embedding technique is performing quite better in our work.

REFERENCES

[1] Deepu, S., Pethuru Raj, and S. Rajaraajeswari. "A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction." Proceedings of the 1st International Conference on Innovations in Computing Networking (ICICN16), Bangalore, India. 2016.

[2] Uchida, Shuto, Tomohiro Yoshikawa, and Takeshi Furuhashi. "Application of output embedding on Word2Vec." 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS). IEEE, 2018.

[3] Rekabsaz, Navid, et al. "Toward incorporation of relevant documents in word2vec." arXiv preprint arXiv:1707.06598 (2017).

[4] Van Nguyen, Thanh, et al. "Combining word2vec with revised vector space model for better code retrieval." 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C). IEEE, 2017.

[5] Amer, Nawal Ould, Philippe Mulhem, and Mathias Géry. "Toward word embedding for personalized information retrieval." Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval. 2016.

[6] Ogheneovo, E. E., and R. B. Japheth. "Application of vector space model to query ranking and information retrieval." International Journal of Advanced Research in Computer Science and Software Engineering 6.5 (2016).

[7] Chen, Dawn, Joshua C. Peterson, and Thomas L. Griffiths. "Evaluating vector-space models of analogy." arXiv preprint arXiv:1705.04416 (2017).