QUERY EXPANSION USING WORD EMBEDDING, ONTOLOGY, AND NATURAL LANGUAGE PROCESSING

Abstract

Query Expansion (QE) is the art of reconstructing specific queries to expand validation presentation, especially in the data mining process in a requirements-understanding environment. Expanding requirements is one of the techniques involved in finding information. In the search engine environment, the query extension includes the evaluation of the value of the construction and the extension of search queries to match new documents. In natural language processing (NLP), word embedding is a term used in textbook parsing, usually as a real-valued vector that encodes the meaning of adjacent words in the vector.It is assumed that the space will be analogous in meaning. Word embedding can be achieved using a set of language models and point literacy methods where vocabulary words or expressions are mapped to vectors of real numbers. For query expansion, one method used is natural language processing through word embedding. Other approaches are ontology, machine learning, and deep learning for automatic query expansion. This chapter proposes a hybridapproach for query expansion by combining NLP and ontology through word embedding.

Keywords: Query expansion, word embedding, natural language processing, ontology, information retrieval.

Authors

Dr. Madhura K
Assistant Professor-Senior Scale
MIT, MAHE Bengaluru
Karnataka, India.
maddyksd87@gmail.com

Dr. Manjula H M
Assistant Professor
School of CSE & IS
Presidency University
Bangalore, Karnataka, India.
manjulahm@presidencyuniversity.in

I. INTRODUCTION

The introduction section of the article introduces the underlying technologies used in the proposed work namely. The technologies used in the proposed work are word embedding through NLP and ontology for query expansion. The article proposes a 3 step procedure for the query expansion. First level, using NLP and second level using ontology and third level by word embedding. Word embedding is one of the best-known expressions in record terminology. It has the ability to understand the placement of words in the report, their semantic and syntactic similarities, their relationships with different words, and more. In natural language processing (NLP), word embedding is the term used to describe words for the purposes of text analysis, and are usually true encodings of word meanings until they are semantically equivalent in a vector space. is used as an estimation vector for [1]. Word embedding can be achieved using a number of linguistic representations and word highlighting strategies that convert jargon words and phrases into vectors of real numbers.
The need for data is one of the principal inspirations for an individual utilizing a web crawler. Questions can address totally different data needs. Unexpectedly, an inquiry can be an unfortunate portrayal of the data need on the grounds that the client can find it hard to communicate the data need. Inquiry Expansion (QE) is by and large prominently used to address this restriction. While QE can be measured as a language-free method, ongoing discoveries have shown that in specific cases, language assumes a significant part [1]. QE can be used to retrieve information from documents, search engines using any category of language. Archive looking through utilizing inquiries that can comprehend the setting can influence the aim and reason for the client's longing while looking through records. Many examinations have been led on understanding the setting of the question, yet contrasts as far as language can prompt various techniques for setting understanding; hence, techniques executed in the past examinations should be moved along [2]. Looking for reports performed by a client requires watchwords to assist clients with acquiring the archive. Watchword addresses the info inquiry from the client in looking through the archive. In any case, the client experiences issues in communicating the inquiry into a significant question. Insufficiency of data recovery frameworks are much of the time brought about by the question in correctness. Countless immaterial reports will be returned if they chose watch words are excessively broad. Recover data from the web utilizing a data recovery framework frequently requires exact watchwords from different field to accomplish the best outcome [3]. The client needs extra inquiry from the framework to assist client with acquiring the pertinent reports. One technique that should be possible is question development. Question development is a strategy for broadening the question term by adding various inquiry possibility to further develop execution in report search.

As a rule, programmed inquiry extension is separated into three: measurable, semantics and mixture [5]. Measurable technique is a question development that investigates inquiries in view of the construction of a given word client in looking. Semantics is a strategy for question development by investigating and understanding the significance of inquiry given by client to acquire the pertinent archive. Crossover is mix technique for measurable what's more, semantics strategy. Crossover is a strategy that is frequently carried out on the grounds that it creates better brings about data recovery [4]. Half breed strategy joins a few inquiry development techniques. The utilization of mixture question development techniques is expected to assist with figuring out the specific circumstance of the inquiry. Understanding

settings on report search is likewise expected to help clients acquire pertinent record results [6].

The ideas in the philosophy can be utilized for word sense disambiguation and ensuing question extension. Question extension has been effective partially yet there is still degree to work on the procedures, connection points or calculations used to gather setting all the more precisely to further develop the outcomes considerably further. Issues to address in question extension incorporate planning calculations for ideal boundary decision; strategies to distinguish when to grow; and managing report assortments which don't have a controlled jargon and are not in every case composed, for example, site pages [7].

Ontologies play an important part in semantic web search. An uncommon use of ontologies in the Semantic Web is to complement existing web resources with some defined tools to improve the search power of today's search engines. The query expansion strategy described in this article is primarily based on each web page and geographic ontology. The proposed strategy differs from traditional strategies as the requirements are strengthened by considering the geographical footprint. In particular, this strategy is designed to fix queries (including castles near Edinburgh) that contain spatial clauses (such as Edinburgh) and complex spatial relationships involving spatial clauses (such as Neighbourhoods). Various factors are believed to help appropriately extend the range question, including spatial terms encoded in geo-ontologies, non-spatial terms encoded in regional ontologies, and the semantics of spatial relationships and usage contexts [8].

Word embedding strategies have received a variety of interest from herbal language processing researchers these days and they're precious aid in figuring out a listing of semantically associated phrases for a seek question. These associated phrases construct an herbal addition for question expansion, however would possibly mismatch whilst the software domain names use different jargon [9].

The objective this paper is to perform the expansion of query using a multi model technology by utilizing word embedding, ontology, and natural language Processing. Recently, word embedding strategy has been widely used to expand queries. The framework is mainly based on the neural language [11]. Based on the internal background, using the K nearest neighbour approach, it extracts sentences that are similar to the questions asked by consumers. Tests are performed on TREC ad-hoc trend data. It confirmed the extended advanced final result compared to the traditional full search approach based on overlapping time periods. Similarly, the word2vec framework primarily based entirely on the extension question was found to do extra or much less the same with uncommented information [10].
The coming section of the article details on the related works carried on query expansion, ontology, NLP and word embedding.

II. REVIEW OF LITERATURE

The second section of the paper covers a deep literature review work carried on ontology, NLP and word embedding for query expansion work.

1. **Review work on Query expansion:** Query growth is a procedure of reformulating a seed question to enhance retrieval overall performance in facts retrieval operations. There are pretty some of researchers firmly agree with that inaccuracy of the question fashioned through some key-word version that the real person facts want is the principle purpose for the ineffectiveness of facts retrieval systems. The predominant motivation of question growth is to feature significant phrases so one can assist the person to put off the ambiguity of the herbal language and additionally specific the facts idea in an extra specified manner into authentic question. By including associated phrases to the authentic question also can growth the wide variety of applicable files identified, for this reason [12].

   Above the program period, where research and studies are done with short questions, a very small amount of knowledge is obtained and additional satisfactory results are obtained. In the 1990s, search engines were introduced and suddenly a lot of knowledge began to be published on the Internet, which has continued to grow at an exponential rate since then. However, users continue to download short queries for web browsing. speed of recall suddenly increases, accuracy is lost (Harman, 1992, Salton & Buckley, 1990). This requires the modernization of the QE method for converting Internet data [13].

   According to the latest report (Keyword, 2018, Statista), the most frequently used queries contain only 1, 2, or 3 words (see Figure 1) - similar to seventeen years ago as reported by Lau and Horwitz (1999). The number of sites is growing fast, even if the search terms are few. This increased anomaly caused by some meaning / sense of the query term (also known as keyword pair problem) to find the relevant page. Therefore, the importance of the QE method in overcoming the lack of words has increased [13].

   However, the QE strategy also has disadvantages, for example, there are computational costs associated with the benefits of the QE strategy. The computational cost associated with the utility of the QE strategy prohibits its partial or complete use when performing Internet searches where short reaction times are important. Another drawback is that it sometimes fails to create misunderstandings between phrases such as "senior citizens" and "elderly" in unfamiliar communities in the corpus. Another problem is that QE can further harm research efficiency for some studies [13].

2. **Review work on NLP in Query Expansion:** Today, a huge amount of digital data is stored. A repository queried by a search system based on a keyword-based interface. Therefore, look for information from the repository. It has become an important issue. Organizations typically implement syntax-agnostic relational database architectures. But most organizations don't do this. Migration due to lack of time, money and knowledge. In this article, we showed you how to perform automation. Query expansion based on natural language processing and semantics. Improved data retrieval from relational database repositories and integration with Real Media Group's existing systems. Organizations have tested it and analysed its effectiveness [14].

   [19] Orland et al. described an approach that considers conceptual semantic theory, using a knowledge base ofconcept networks to extend queries. A new query is derived from the conceptnetwork corresponding to the query term. Semantic similarities

between concepts are formulated and represented using a directed graph model called Conceptual Word Cluster Space Graph (CWCSG). Here, the user's query is augmented by herto exactly meet the user's search needs. Compared to the WordNet [20]synonym dictionary, the above method has better performance, but the quality of the concept network is a very important consideration. Based on a threshold, terms with higher weights are derived, and based on these top n documents, [21] related terms are selected. Query expansion by G. Akrivas et al. It only added semantic entities, no terms were considered.

Princeton University has proposed WordNet, an online framework for philological situations. Verbs, nouns, adjectives and adverbs are grouped into substitution groups (synsets). Synsets are controlled by faculty. Synsets are further linked to synsets classified according to specific categories of matter. Most common compound words are hyponyms/hypernyms (ie, Is-A complexes) andmeronyms/homonyms (ie, parts of complexes). There are nine of them and some verb-is-a hierarchies (adjectives and adverbs are unordered in the is-a hierarchy) [22].

3. **Review work on ontology in Query Expansion:** Due to the large amount of data emerging on the web, traditional keyword searches are being challenged by short queries that users submit to explain a vague need for information. Query Extensions have been studied for decades, and various extension strategies have improved search performance. Currently, the approach to expanding the query is based on knowledge as the web becomes more meaningful, it becomes more popular. This paper examines the cutting edge of ontology-based query expansion approaches and develops practical strategies for leveraging rich ones. Ontology Semantics The scope of this thesis, on the one hand, focuses on finding success factors for ontology-based query extensions. On the other hand, it emphasizes exchange between the search efficiency obtained and the computational cost incurred [15].

In 2021, Namrata Rastogi et al., proposed a work on query expansion based on word embedding plus ontology. The work was proposed for information retrieval [16].

In 2019, Maryamah Maryamah et al., proposed a research work on query expansion for searching the Arabic documents using wikipedia word embedding and babelnet method [17].

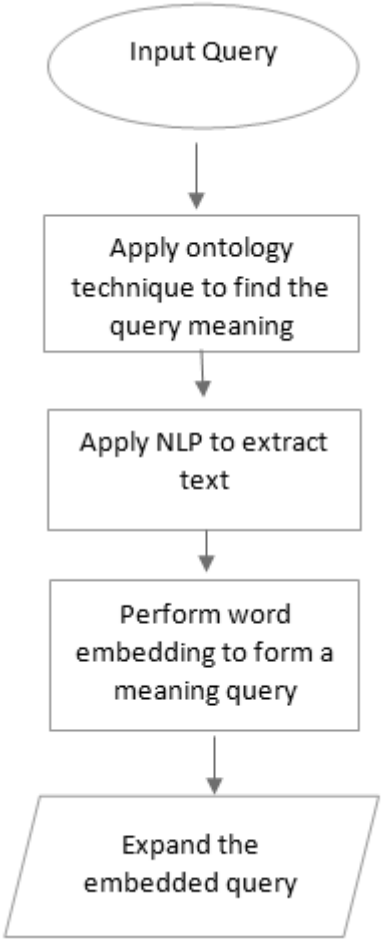
In 2020, Doguhan Yeke submitted a thesis on Improving Document Ranking with Query Expansion Based On Bert Word Embedding. Bert word embedding utilized fine tuning and feature based approaches [18].

III. PROPOSED WORK

Proposed work is a hybrid approach which combines technologies like ontology, NLP and word embedding for query expansion.

1. **The steps involved is listed below:**

Step 1: Apply ontology to understand the meaning of the query.
Step 2: Apply NLP technique to extract the text contents from the query.
Step 3: Finally perform word embedding to form a meaningful query.
Step 4: Retrieve the embedded query from step 3 and form the final expanded query.



2. **Algorithm:**

Input: Query Q

Repeat until query meaning is true

Perform feature extraction

If Feature Extraction is true; then

Perform word embedding till EOF

Output: Retrieve the output query Q1

3. Explanation of the Algorithmic step:

- Give the input query, for example, lets us say a Query Q.
- Do the feature extraction of the query until the meaning of the query is found.
- Once the features are extracted, the next process is to embed the extracted featured words till the last character.
- If the formed words are meaningful then can be treated as a newly formed query.
- This formed newly formed query is the output query Q1.

IV. CONCLUSION

In the existing work on the query expansion uses the technologies like NLP and ontology as separate techniques. In the present works, the hybrid framework for query expansion is present in terms of either NLP plus word embedding or ontology plus word embedding.

As the efficiency of hybrid approach is more compared to the individual apply of techniques, the proposed work combines all three techniques together in terms of ontology, NLP and word embedding for query expansion.

REFERENCES

[1] Farhan YH, et al., (2020). Survey of Automatic Query Expansion for Arabic Text Retrieval. Journal of Information Science Theory and Practice, 8 (4), 67–86. https://doi.org/10.1633/JISTAP.2020.8.4.6.
[2] Maryamah Maryamah, et al., “Query Expansion Based on Wikipedia Word Embedding and BabelNet Method for Searching Arabic Documents”, International Journal of Intelligent Engineering and Systems, Vol.12, No.5, 2019 DOI: 10.22266/ijies2019.1031.20.
[3] J. Ooi, et al., “A survey of query expansion, query suggestion and query refinement techniques”, In: Proc. Of 2015 4th Int. Conf. Softw. Eng. Comput. Syst. ICSECS 2015 Virtuous Softw. Solut. Big Data, pp. 112–117, 2015.
[4] M. A. Raza, et al., “A survey of statistical approaches for query expansion”, Knowl. Inf. Syst., 2018.
[5] B. El Ghali, et al., “Context-aware query expansion method using Language Models and Latent Semantic Analyses”, Knowl. Inf. Syst., Vol. 50, No. 3, pp. 751–762, 2017.
[6] J. Bhogal, et al., “A review of ontology based query expansion”, Information Processing & Management, Volume 43, Issue 4, 2007, Pages 866-886, ISSN 0306-4573,
[7] https://doi.org/10.1016/j.ipm.2006.09.003.
[8] Fu, G., Jones, C.B., Abdelmoty, A.I. (2005). Ontology-Based Spatial Query Expansion in Information Retrieval. In: Meersman, R., Tari, Z. (eds) On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE. OTM 2005. Lecture Notes in Computer Science, vol 3761. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11575801_33.
[9] Vuong M. Ngo, Tru H. Cao (2018), “Ontology-Based Query Expansion with Latently Related Named Entities for Semantic Text Search”, accepted by Advances in Intelligent Information and Database Systems, Book of series SCI, Springer-Verlag, https://doi.org/10.48550/arXiv.1807.05579.
[10] Andre Rattinger, Jean-Marie Le Goff, and Christian Guetl, Local Word Embeddings for Query Expansion based on Co-Authorship and Citations, BIR 2018 Workshop on Bibliometric-enhanced Information Retrieval.
[11] Tarun Goyal | Ms. Shalini Bhadola | Ms. Kirti Bhatia "Automatic Query Expansion Using Word Embedding Based on Fuzzy Graph Connectivity Measures" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456- 6470, Volume-5 | Issue-5, August 2021, pp.1429-1442, URL: www.ijtsrd.com/papers/ijtsrd45074.pdf
[12] Roy, D., Paul, D., Mitra, M., & Garain, Using word embeddings for automatic query expansion, arXiv preprintarXiv:1606.07608, 2016.

[13] J. Ooi, Xiuqin Ma, Hongwu Qin and S. C. Liew, "A survey of query expansion, query suggestion and query refinement techniques," 2015 4th International Conference on Software Engineering and Computer Systems (ICSECS), 2015, pp. 112-117, doi: 10.1109/ICSECS.2015.7333094.
[14] Azad, Hiteshwar Kumar, Deepak, Akshay, Information Processing & Management,VL - 56, IS - 5, SP - 1698, EP - 1735, PY - 2019,SN - 0306-4573, DO - https://doi.org/10.1016/j.ipm.2019.05.009, UR - https://www.sciencedirect.com/science/article/pii/S0306457318305466.
[15] Mar´ıa G. Buey, Angel Luis Garrido, and Sergio Ilarri, "An Approach for Automatic Query Expansion Based on NLP and Semantics", H. Decker et al. (Eds.): DEXA 2014, Part II, LNCS 8645, pp. 349–356, Springer International Publishing Switzerland 2014.
[16] Jiewen Wu, Ihab Ilyas, and Grant Weddell, "A Study of Ontology-based Query Expansion", Technical Report CS-2011-04.
[17] Namrata Rastogi et al., "Query Expansion based on Word Embeddings and Ontologies for Efficient Information Retrieval", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 11, 2021.
[18] Maryamah Maryamah et al., "Query Expansion Based on Wikipedia Word Embedding and BabelNet Method for Searching Arabic Documents", International Journal of Intelligent Engineering and Systems, Vol.12, No.5, 2019, DOI: 10.22266/ijies2019.1031.20.
[19] Doguhan Yeke et al., "IMPROVING DOCUMENT RANKING WITH QUERY EXPANSION BASED ON
[20] BERT WORD EMBEDDINGS", A THESIS SUBMITTED TOTHE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCESOFMIDDLE EAST TECHNICAL UNIVERSITY in July 2020.
[21] Hoeber, O., Yang, X.D. and Yao, Y. Conceptual query expansion. International Conference on Atlantic Web Intelligence, 2005, 190-196.
[22] Peng, M., Lin, Q., Tian, Y., Yang, M., Xiao, Y. and Ni, B. Query expansion based on Conceptual Word Cluster Space Graph. IEEE 5th International Conference on New Trends in Information Science and Service Science (NISS), 2011, 128-133.
[23] Jain, A., Mittal, K. and Sabharwal, S. Conceptual weighing Query Expansion on user profiles. National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC Proceedings published by International Journal of Computer Applications (IJCA), 2012.
[24] Mittal, K. and Jain, A. Word Sense Disambiguation Method using Semantic Similarity Measures and OWA Operator. ICTACT Journal on Soft Computing 5 (2) (2015) 896-904.

