People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Constantine 2 University – AbdelHamid Mehri

Faculty of New Information and Communication Technologies, Department of Software Technologies and Information Systems

Option :Data Science and Intelligent Systems

QueryExpander : Enhancing Information Retrieval with NLP and Word Embeddings

---

Made by :
Menaifi Maya Fairouz
Ouchene Aya

Under the supervision of :
Dr. Aklouche Billel

7 avril 2024

Table des matières

Chapitre 1

Enhancing Information Retrieval through Query Expansion :

1.1 Summary :

1.1.1 Introduction :

The paper titled "Query Expansion Based on NLP and Word Embeddings" discusses various techniques aimed at enhancing search queries to improve the relevance and richness of search results. Query expansion methods leveraging Natural Language Processing (NLP) tools and Word2Vec embeddings are explored in detail. Specifically, the paper delves into the use of Word2Vec models, namely Skip-Gram and Continuous Bag Of Words (CBOW), to generate word embeddings that represent semantic relationships between words.

1.1.2 Methodology Overview :

1.1.2.1 Data Pre-processing :

The proposed approach begins with data pre-processing, involving the conversion of the corpus into a clean and consistent format. This process includes text segmentation, sentence tokenization, and stop words removal to filter out non-informative words. Additionally, stemming is applied to reduce all terms to their respective stems, ensuring uniformity in word representations.

1.1.2.2 Model Training :

Following data pre-processing, the Word2Vec model is trained using either the Skip-Gram or CBOW architecture. Parameters such as window size and vector dimensions are set, and words appearing less than 5 times in the corpus are removed. The trained model computes word embeddings, generating vector representations of words based on their context within the corpus.

1.1.2.3 Query Expansion :

In the query expansion phase, word vectors are utilized to identify semantically related terms to the original query. Euclidean Distance is employed to calculate the similarity score between vectors, enabling the selection of expansion terms that are similar to the entire query or its individual terms. The chosen expansion terms are sorted based on their similarity scores, with the top n terms selected to augment the original query.

1.1.2.4 Document Ranking :

The expanded queries are then used in document retrieval and ranking, where the Terrier4 Information Retrieval platform, along with the Okapi BM25 weighting scheme, is employed to index collections and rank documents. The performance of various runs is evaluated based on existing relevance judgments provided by NIST assessors.

1.1.3 Results and Discussion :

The paper presents the results of 72 runs submitted, with a comparison of manual, existing relevance judgments, and automatic evaluations. The best-performing run, jarirsgre, utilizes the Word2Vec Skip-Gram model,

expansion terms similar to the entire query, and query reweighting. Notably, this approach consistently outperforms others, with query reweighting demonstrating significant improvement across all evaluation metrics. Additionally, the Word2Vec Skip-Gram model generally outperforms the CBOW model, although the latter yields slightly better results when selecting expansion terms similar to individual query terms.

1.1.4 Conclusion :

The paper highlights the effectiveness of utilizing NLP techniques and Word2Vec embeddings for query expansion, resulting in improved information retrieval performance. The study underscores the importance of preprocessing data, training robust models, and strategically expanding queries to enhance search relevance. The findings suggest that leveraging the Skip-Gram architecture and selecting expansion terms similar to the entire query can lead to superior retrieval outcomes, with query reweighting further augmenting performance.