I.J. Information Technology and Computer Science, **2023, 1, 42-55**
Published Online on February 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2023.01.05

A New Query Expansion Approach for Improving Web Search Ranking

Stephen Akuma*
Institution: Department of Mathematics and Computer Science, Benue State University
E-mail: sakuma@bsum.edu.ng
ORCID iD: https://orcid.org/0000-0003-1909-7618
*Corresponding author

Promise Anendah
Institution: Department of Mathematics and Computer Science, Benue State University
ORCID iD: https://orcid.org/0000-0002-8600-2407
E-mail: anendahpromise@gmail.com

Received: 23 August 2022; Revised: 02 October 2022; Accepted: 11 November 2022; Published: 08 February 2023

Abstract: Information systems have come a long way in the 21st century, with search engines emerging as the most popular and well-known retrieval systems. Several techniques have been used by researchers to improve the retrieval of relevant results from search engines. One of the approaches employed for improving relevant feedback of a retrieval system is Query Expansion (QE). The challenge associated with this technique is how to select the most relevant terms for the expansion. In this research work, we propose a query expansion technique based on Azak & Deepak's WWQE model. Our extended WWQE technique adopts Candidate Expansion Terms selection with the use of in-links and out-links. The top two relevant Wikipedia articles from the user's initial search were found using a custom search engine over Wikipedia. Following that, we ranked further Wikipedia articles that are semantically connected to the top two Wikipedia articles based on cosine similarity using TF-IDF Vectorizer. The expansion terms were then taken from the top 5 document titles. The results of the evaluation of our methodology utilizing TREC query topics (126-175) revealed that the system with extended features gave ranked results that were 11% better than those from the system with unexpanded queries.

Index Terms: Search Engine, Query Expansion, Relevance Feedback, Information Retrieval, WWQE Model.

1. Introduction

Information retrieval systems such as search engines are becoming more relevant with technological advancement. Processing the vast amounts of unstructured data found on the internet is extremely difficult because most internet users are unfamiliar with search strategies [1, 2]. Traditionally, search engines work on a simple principle; get the users' queries, search and return the best results. The feedback users get from search engines lies more in their ability to construct good queries with the right keywords that apply to the search, than in the amount of available information [6-9]. So, if the input is not well constructed or structured, it can lead to poor results regardless of the optimizations that might go into the search and ranking algorithms of search engines [3-5].

To convey their problem statement in a search engine, users who are looking for the same concept typically use different vocabulary. For instance, comparable phrases like "risky," "chancy," and "peril" may be used by searchers as query terms, but only one of these terms may be indexed in a page. The retrieval of relevant results for search terms with the same semantic meaning but no index in documents may be affected by this [6]. Although natural languages have been used in search engines to improve Search Engine Result Pages (SERP), it is still challenging to capture users' intent to return relevant results from inputted queries[7]. Some Search engines employ the use of query suggestions [8] but they are still unable to help novice searchers effectively represent their intent. This problem often causes users to open many documents or carry out optimistic query reformulation to arrive at the desired results, leading to prolonged search sessions and frustration [9]. Even though this is not an issue for domain experts [4], it remains a challenge for most internet users.

To improve information retrieval systems, different techniques have been experimented upon by researchers [1, 10, 11]. Query Expansion (QE) as a way of reformulating queries and improving retrieval results has received a lot of attention from researchers[1, 12]. It is a technique of finding suitable terms for the reformulation of queries to resolve the short query and word mismatch problem and improve the performance of information retrieval systems [13]. Unlike query refinement which changes the original query by complete removal of terms, query expansion focuses on adding meaningful terms to reduce the ambiguity of the communication language while expressing the search in a more detailed way [14]. The process of adding more terms into a query could be automatic or manual [15]. Manual query expansion relies on the user's input to decide the terms that will be added to the query, this could be mentally strenuous for users (especially novice users) and may not be helpful due to a lack of vocabulary familiarity. Automatic Query Expansion (AQE) on the other hand uses term weighting to add terms that will produce useful results and reduce mental stress [16, 17].

Considerable work has been done in understanding users' intent from queries and optimization of query expansion techniques [1, 12] but little focus has been given to the incorporation of query expansion techniques into modern search engines and selecting the best terms to incorporate. When relevant terms are added to the queries, information retrieval effectiveness is enhanced [18]. Although different data sources for QE like thesaurus, ConceptNet, WordNet, Corpus, Web, Wikipedia and other hybrid data sources[18] have been used by researchers, they are, however, not effective in a non-domain specific environment and in matching semantically related words. Previous research has shown that Wikipedia and WordNet data sources are often employed as knowledge resources to enhance the semantics of the initial query during QE [18-21].

The goal of this work is to use a novel query expansion technique to enhance the feedback (SERPs) of contemporary search engines. To achieve this, we put forth a query expansion strategy based on [18] with enhancements to document weighting and retrieval of Candidate Expansion Terms (CETs). The following are this paper's main contributions: 1) Real-time Query Expansion (RTQE) model development based on Wikipedia; 2) Document weighing and CET improvement; and 3) Evaluation of the proposed query expansion model using the TREC dataset. The rest of the paper is organized into the following: Section 2 presents related work summarizing past research on query expansion theories and methodologies; Section 3 focuses on a detailed description of our approach; Section 4 presents the result of the implementation; the evaluation is presented in section 5, while the conclusion and future work are presented in Section 6.

2. Related Works

This section looks at some of the concepts of information retrieval and elaborates on past research related to this study in theory and ideas with key attention to the most recent research. The most common solution proposed over the years to improve the effectiveness of information retrieval systems is query modification techniques such as query expansion and query refinement [16]. Ooi et al. [16] discussed these techniques and made a comparison to distinguish them. Bisht & Bisht [22] researched to find the effect of query formulation on web search engine feedback. In their study, the Google search engine was used. They examined the number of common documents from sample queries that were semantically the same but structurally different. The result of their experiment shows that there is no significant difference in the total number of documents from the different queries except for the first five and first ten documents. Since most users hardly view more than the first two results on the SERP [23], users may get different results from different query representations. However, QE approaches can be used to expand such structural different queries to return relevant results.

Research has found automatic query expansion useful. Carpineto et al. [15] surveyed automatic query expansion and found that AQE has gained much popularity due to reports at TREC of noticeable improvements in retrieval performance by participants who make use of the technique. The work of Claudio & Giovanni [24] noted that with AQE, there is more chance of retrieving documents that don't have the original terms. The new query does not only get documents with the original terms but also documents that use different spellings. Kucukyilmaz [12] posits that no action is required by the user on how the expansion should be done thereby leaving the user to focus on the task at hand. Sharma, Dilip Kumar, Pamula et al. [25] say that QE reduces significantly the number of redundant results returned by search engines. However, other researchers argue that AQE techniques are computationally expensive and will negatively impact the latency of search engines on implementation [26]. Xiong and Callan [27] added that automatic query expansion often damages many queries, making it risky for online search services as users are more sensitive to failures than successes [27]. There are also concerns regarding the acceptance of automatic query expansion due to the limited transparency of its implementation in IR systems [15, 28]. However, the effectiveness of QE usually lies in the techniques used in implementing it. Researchers have used different techniques to implement automatic query expansion over the years, from the use of term distribution, relevance feedback and fuzzy logic to word embeddings, random walk models and freebase [25]. In section 2.1, we review some of the various techniques used in the implementation of query expansion.

2.1. Query Expansion Techniques

Yonggang and Frei [29] used a probabilistic query expansion model based on a similar thesaurus and the domain knowledge of the search. Years after Yonggang and Frei's [29] research, Hang Cui et al. at Microsoft Research Asia, proposed and experimented with a probabilistic approach with query log mining, to improve search performance [30].

Fonseca et al. [31] used a different query expansion framework that involved mining, identifying and labelling query relations. Contrary to previous research that used the term correlation technique, Riezler et al. [32] used a modified approach (contextual query expansion) that translated queries into snippets. Pal et al. [19] worked on the use of 'Term Distribution and Term Association' in query expansion; Xiong and Callan [27] used Freebase in their work to improve the performance of query expansion. Recently, Word Embeddings have been used which encompasses Wikipedia and WordNet [18, 20].

Pal et al. [19] proposed a new way of using WordNet to improve query expansion. The work used the relevance feedback technique with the innovation of using WordNet to measure the usefulness of candidate expansion terms found in the pseudo-relevant documents. They proposed three methods of determining the usefulness of candidate query terms which are as follows; a distribution-based method, an association-based method and a WordNet approach. According to Pal et al. [19], this new approach out bested methods like KLD and RM3 over standard TREC collections. The work also proposed a combination of methods such as term distribution and association in the target corpus (WordNet) and finding the semantic relationship of expansion terms with query terms to decide the usefulness of expansion candidate terms. Their method was tested on the TREC dataset and the result shows that the combined approach was better than the individual methods. Keikha et al. [33] used Wikipedia data sources for query expansion and they showed that Wikipedia is more effective than WordNet in expanding queries for information seeking and general queries.

Freebase dataset was also used for query expansion. Xiong and Callan [27] focused on the improvement of query expansion with Freebase, a large knowledge base which contained 2.9 billion relationships and attributes and about 48 million topics. Their solution was decomposed into two components: The first component identified the topics to be used for expansion while the second component used information about the topics to select candidate terms for query expansion. They experimented using ClueWeb09, TREC web Track 2009-2012 with relevance judgement by TREC annotators, and it yielded an almost 30% gain in effectiveness than some state-of-the-art expansion methods.

Other researchers aggregated the methods highlighted above for query expansion. Azad & Deepak [18] aggregated WordNet and Wikipedia for AQE improvement. They used Wikipedia data sources to generate phrase terms and WordNet data sources to provide expansion terms for individual queries. This method also referred to as the Wikipedia-WordNet-based query expansion technique (WWQE), showed gains of about 24% on the Mean Average Precision (MAP) score and about 48% on the Geometric Mean Average Precision (GMAP) score over expanded queries on the FIRE dataset, revealing its effectiveness. Azad & Deepak [14] extended the research behind the WWQE model, proposing the use of pseudo-relevant web knowledge, consisting of the top N web pages returned (in response to the original query) by three popular search engines (i.e. Google, Bing and DuckDuckGo). These documents were used with three weighting models; TF-IDF, K-Nearest Neighbour (KNN) based on cosine similarity and correlation score. Their proposed model called Web Knowledge-based query expansion (WKQE) achieved an improvement of 25.89% on the MAP score and 30.83% in the GMAP score over unexpanded queries on the FIRE dataset. These results when compared to their previous implementation [14] reveal that this method outperformed the first approach by 1.85% when used on similar queries while its performance was poorer on a wide range of different kinds of queries revealed by an 18.83% drop on the GMAP score.

Even though AQE has been shown by previous researchers to be very effective, concerns have been raised about the effectiveness of query expansion in search engines by other researchers [15, 24, 28]. Thus, this research investigates this concern to ascertain whether query expansion returns more relevant results to users. The successes/limitations of AQE are further investigated in this work through a new expansion methodology.

3. Methodology

To improve search results by query expansion, a query expansion model is required. The model must successfully expand search queries with significant performance improvement. In this research work, the WWQE (Wikipedia WordNet Query Expansion) methodology by Azad and Deepak [18] has been adopted for the development of the query expansion model with significant improvement in the aspect of Candidate Expansion Terms (CET) retrieval and document weighting. Their method used a combination of WordNet and Wikipedia expansion techniques for query expansion. Azad & Deepak [18] method was chosen to form a foundation for this research because their expansion method improved the quality of results as compared to other query expansion techniques. Also, their data source particularly Wikipedia is a large body of organised information covering a very wide range of topics with updates every 2 seconds from contributors around the world (Wikipedia: Statistics), making it the ideal data source for this work. The minimum system requirements to run our expansion software is 2MB/S internet connectivity speed, 2.5GHz processing power with multi-threading capabilities for parallel computing and Python 3.9 Interpreter. The architecture of the proposed system is presented in Fig. 1.

Our extended WWQE technique adopts CET selection with the use of in-links and out-links. We used a custom search engine over Wikipedia to obtain the top 2 related Wikipedia articles from the user's original query. We then crawled all other Wikipedia documents that are semantically related to each top document. The relevance of each semantically related document to the 2 top relevant documents was measured. We then rank the results and use the top 5 document titles as our expansion terms. This method is further explained in detail in Section 3.1

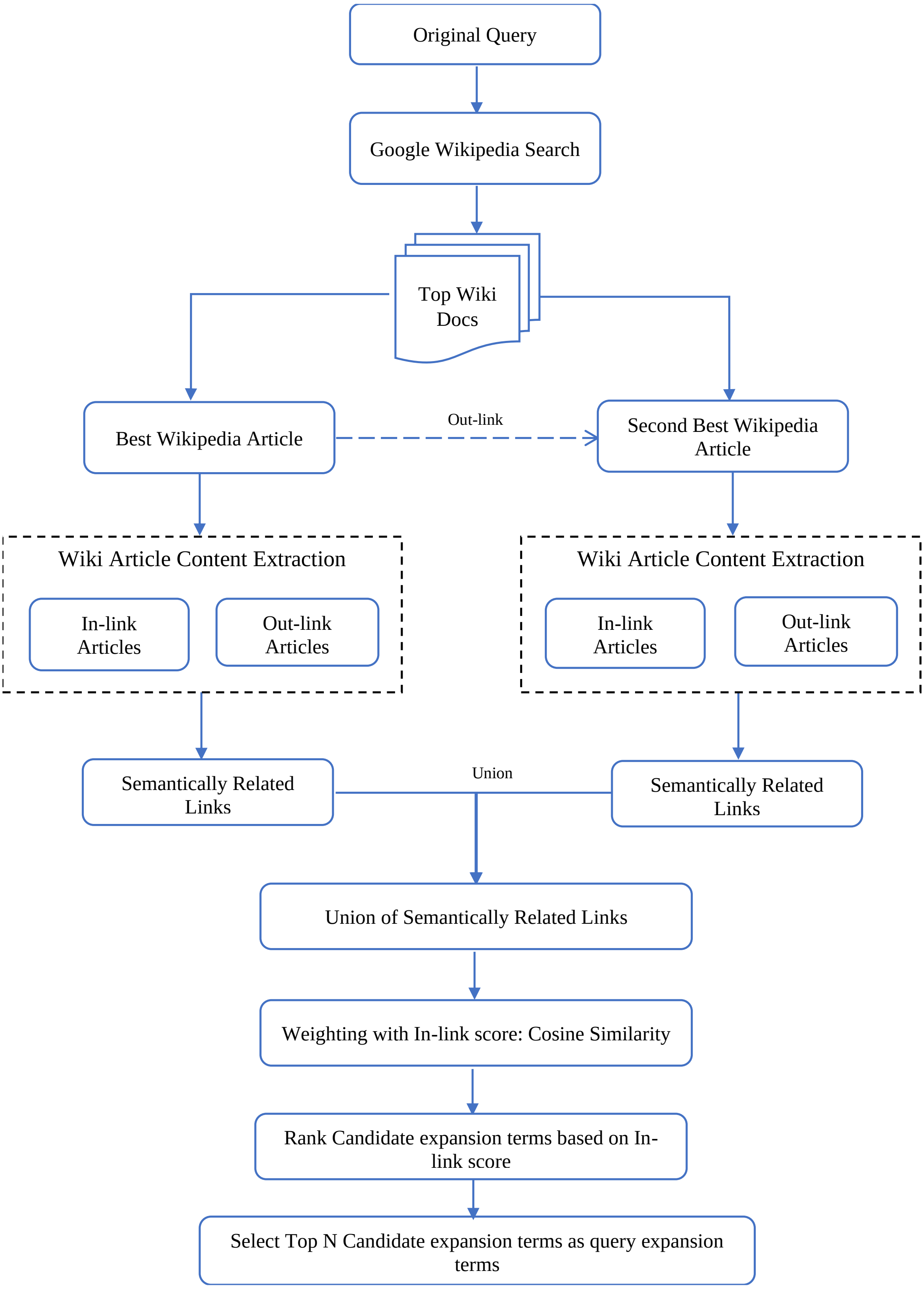

Original Query
Google Wikipedia Search
Top Wiki Docs
Best Wikipedia Article
Out-link
Second Best Wikipedia Article
Wiki Article Content Extraction
In-link Articles
Out-link Articles
Wiki Article Content Extraction
In-link Articles
Out-link Articles
Semantically Related Links
Union
Semantically Related Links
Union of Semantically Related Links
Weighting with In-link score: Cosine Similarity
Rank Candidate expansion terms based on In-link score
Select Top N Candidate expansion terms as query expansion terms

Fig.1. Proposed Approach.

3.1. Query Expansion with Wikipedia

To match the best document to inputted queries, we used Google programmable search engine customized to search only the English Wikipedia domain (https://en.wikipedia.org). The search engine has features such as Google page rank algorithm, Google spelling correction, Google query analysis and other features not used in this research, thereby making the retrieval of the best Wikipedia article from a query possible with better accuracy and time performance. This method worked well for long queries (or queries with a lot of stopwords) and the results were returned in JSON format.

Our approach to query expansion with Wikipedia requires at least one target Wikipedia article that relates to the original query. Unlike the method used by Azad & Deepak [18] where they worked with a Wikipedia dump and used a locally crafted method for matching query phrases and individual terms to Wikipedia articles, we improve this method of article retrieval by utilizing Google page ranking algorithm via a programmable search engine. This involves retrieval of a relevant Wikipedia article, document content extraction, extraction of in-links, extraction of out-links, assignment of the in-link score to expansion terms and selection of top n terms as expansion terms as elucidated in section A to section G.

A. Retrieval of Relevant Wikipedia Articles

We retrieved relevant Wikipedia articles using Google Wikipedia Search programmable search engine by making a request using the format below:

https://customsearch.googleapis.com/customsearch/v1?key=API_KEY&cx=047d9bfb192dc6725&q=searchQuery &start=1&alt=json

The search query in the link above is the user's query unmodified. We then obtained a list of the best 10 results with a spelling correction. From the list of results from Google Wikipedia Engine, we obtain the best article A. We then get the contents of the article as well as its out-links. Next, we get the second result (article B from our search engine as well as its content and out-links), which is also an out-link of the first result, this is to ensure that we build around the same topic without diverging since we aim to expand to different areas of the same topic. These two documents form the foundation for our query expansion. They serve as a reference to other documents on the same topic.

B. Document Content Extraction

The page content of Wikipedia articles involved in query expansion is parsed and extracted. We used the extracted content to calculate the relevance of the article. While the body write-up, out-links and in-links of the 2 reference articles are gotten, only the in-links and body write-up of the other documents are retrieved.

C. Extraction of In-Links

An in-link x1 (Wikipedia article that contains a hyperlink to the query term, the query term here simply refers to a subset of the set P of all phrases and individual words from the original query) is represented by:

$$I(x) - \{x_i \mid (x_i, x_n) \in L\} \tag{1}$$

Where I denote a superset of in-links to an article and L is a superset of all Wikipedia links.

After the extraction of the in-links, the term frequency of the initial query terms is computed for each in-link. This was done by finding the term frequency of the query term t_1 and its synonyms obtained from Wikipedia in the in-link article x_1.

D. Extraction of Out-Links

Out-Links (which are hyperlinks within the body of the article of the query term) were extracted from hyperlinks from the Wikipedia page of the query term and represented thus:

$$O(x) - \{x_i \mid (x_i, x_n) \in L\} \tag{2}$$

Where O denotes a superset of Out-Links to an article and L is a superset of all Wikipedia links.

E. Assignment of In-Link score to expansion terms

After the extraction of the In-Links and the Out-Links of the query term, expansion terms were selected from the out-links on a semantic similarity basis shown in Fig. 2. A query term t and an expansion term t_1 are said to be semantically similar if t_1 is both an in-link and an out-link of t and t_1.

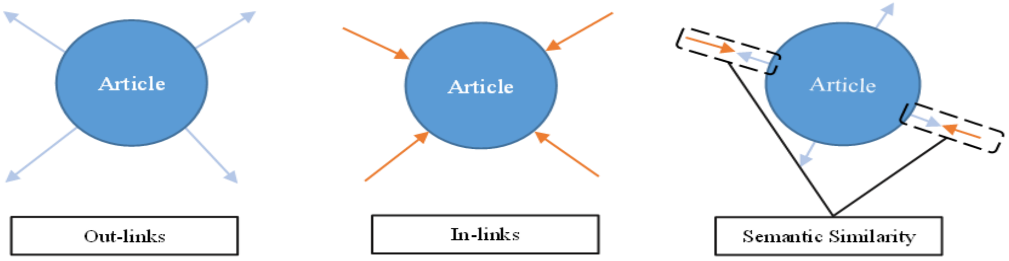


Fig.2. Out-links, In-Links and Identification of semantic similarity.

Our In-link score is based on cosine similarity using TF-IDF Vectorizer [11]. Given two documents A and B their cosine similarity can be defined as:

$$\text{Cosine Similarity} = S_C\,(A, B){:} = \cos(\Theta) = \frac{A\ .\ B}{||A||||B||}\ \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3}$$

Where A_i and B_i are TF-IDF weights of the terms in documents A and B respectively.

Cosine similarity was used in this work to make sure the documents that are semantically related indeed relate to the 2 relevant articles. The higher the cosine of two documents, the higher the similarity. Cosine similarity is the closed interval [-1, 1] where a value of 1 means maximally similar and -1 means maximally dissimilar.

$$\text{tf}(t, D) = \frac{total\ count\ of\ t\ in\ D}{total\ number\ of\ terms\ in\ D} \tag{4}$$

$$\text{idf}\ (t_1, W_D) = \log \frac{N}{\{d \in WD : t1 \in d\}} \tag{5}$$

Where:

tf(t, D) – is the normalized term frequency of the term t in document D

N – is the total number of articles on Wikipedia, and

$|\{d \in W_D : t1 \in d\}|$ – is the total number of Wikipedia articles that contain query terms.

F. Selection of Top N Expansion Terms

At this point, we have obtained the 2 best Articles as reference points as well as their contents and semantically related links. The expansion terms are weighted against the 2 reference documents (combined) using cosine similarity with TF-IDF vectorizer to eliminate document length bias. The weighted terms were then sorted in descending order of weight magnitude with the best terms appearing on top. These top N terms make up our expansion terms that will then be used to expand the query.

The steps listed above were followed by a clean-up operation where symbols and other redundant characters are removed by tokenizing each expansion term. Expansion terms that appear in the original query were then appended to the original query to obtain the expanded query. Our method makes use of parallel computing techniques in both the data retrieval and document pre-processing phases to improve performance (response time). We use asynchronous requests which allow multiple independent requests to be sent at the same time in a batch.

3.2. Data Collection

To obtain the specific data relevant to this research, many APIs were used. We used Wikipedia as our data source for this research work because of the way it is organised, the large volume of data, the currency and the acceptability by researchers. Wikipedia has a large volume of data, as of April 2021 when this research was conducted, the size of all articles compressed was about over 19.52GB including text and multimedia resources which can be downloaded via its WikiDum, however, to develop a reliable system that will be up to date without constantly manually updating the data content, we opted for live data collection of Wikipedia data for this research, by this, we easily made use of data in real-time.

4. Results

We have proposed a query expansion model based on Wikipedia articles. This section discusses the processes involved in the expansion of queries using our proposed method as well as how to improve the performance of the expansion. The query expansion model has been developed using the python programming language with the help of a Google Programmable Search Engine and Live data from Wikipedia. The goal was to integrate this query expansion model into a search to show that this method can improve the relevancy of documents returned by search engines.

4.1. Implementation of Query Expansion Module

The query expansion module was developed based on our proposed expansion model. This expansion software was written in the python programming language. Development was done in different environments. The environments used during the development of this expansion module include Windows 10 64-bit PC, JetBrains Datalore online platform and Google Colab. The shift from one environment to another was due to the changing network and hardware requirements. Development started on a local Windows 10 64-bit PC but was moved to Jetbrains Datalore due to the increasing network connectivity requirement which was necessary to not just make the expansion possible but also to do so in good time. Datalore provided a fast connection to our source of data which made expansions faster. We later switched to Google Colab when we introduced Cosine Similarity in our methodology which required more processing power.

4.2. Deployment of our Expansion Module

We deployed our query expansion software as a separate system fully functional on an independent machine different from the host of the search engine. This decentralization enabled scalability and proper maintenance or configuration changes. We hosted it as an app on the Amazon Web Services cloud infrastructure using the Elastic Beanstalk environment. This is aimed at increasing the computing power and speed of connection to data sources.

4.3. Query Expansion as Web Service

Since our expansion software has been made an internet application through cloud hosting, it provides services (expansion services) to requesting applications using standard communication protocols over HTTPS. We designed a simple RESTFUL API that can allow service requesters (e.g our custom search engine) to request for expansion of queries and get a result. The query expansion service is the only service that our application offers, therefore, our API has only one parameter (query) which must be specified in the request.

4.4. Custom Search Engine Implementation

The custom search engine was developed to limit the search to a specific domain for evaluation. It uses Google's famous page rank algorithm, and it was reconfigured to return results of a specific domain. Fig. 3 and Fig. 4 show the flowchart and architecture of the custom search engine. Although the custom search engine feedback is not the same as that of Google search, the first ten results from its results are similar. This is because the custom engine does not contain all of Google's features.

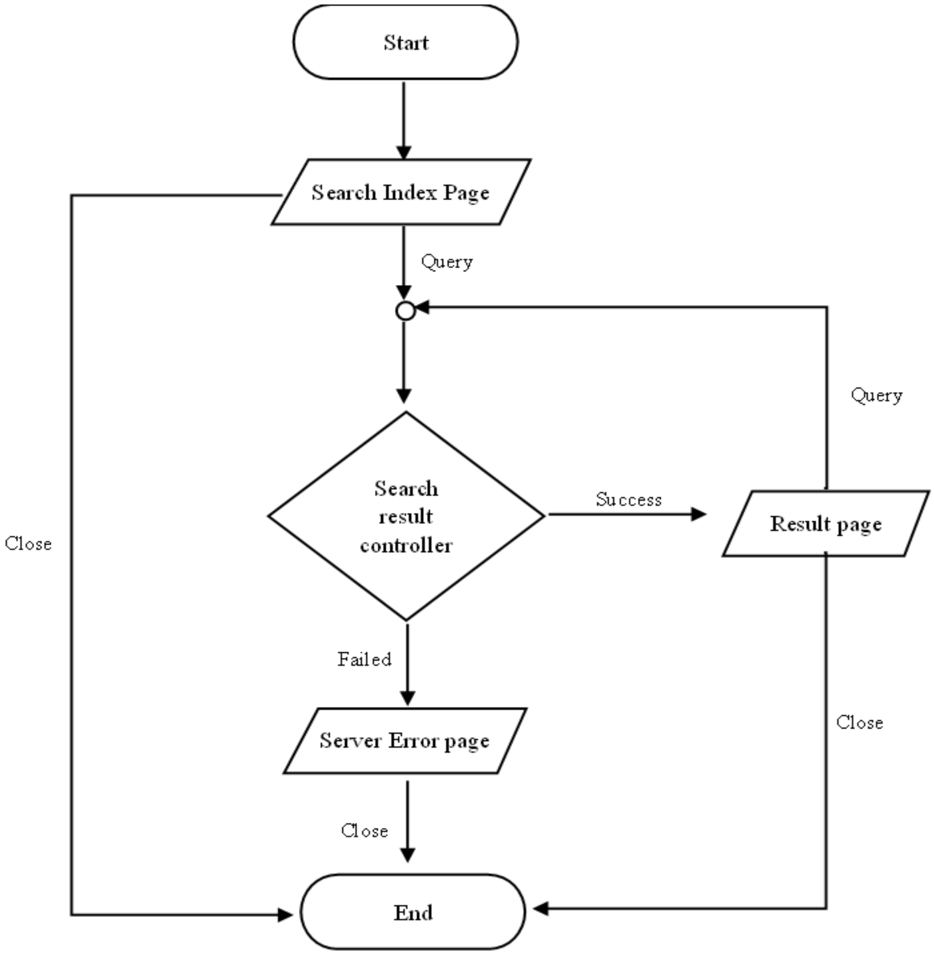


Fig.3. Flowchart of the Custom Search Engine.

Our custom search engine has 2 major components: the user interface (the result page and the error page which serve as the front end) and the Search Engine with the custom search implementation also referred to as the google search request handler and result parsing scripts.

4.5. Integration of Query Expansion Module

In this section, we discuss the integration of our query expansion module into our search engine. Contrary to traditional software integration processes, our expansion function lies on the web as an independent system. We configured our custom search engine to request expansion from our expansion service before proceeding to search the internet with the newly expanded query. Our integration process includes the following steps:

- Implementation of a request handler (coupled with our search engine search result controller) in the custom search engine application that can communicate with our expansion service.
- Parsing the result to the expansion service.
- Request the custom search engine with a modified query.

The design is such that if the query expansion software fails, the search engine will go ahead and make the search with the original query. The flowcharts for the expanded search engine and architecture are depicted in Fig.3 and 4 respectively.

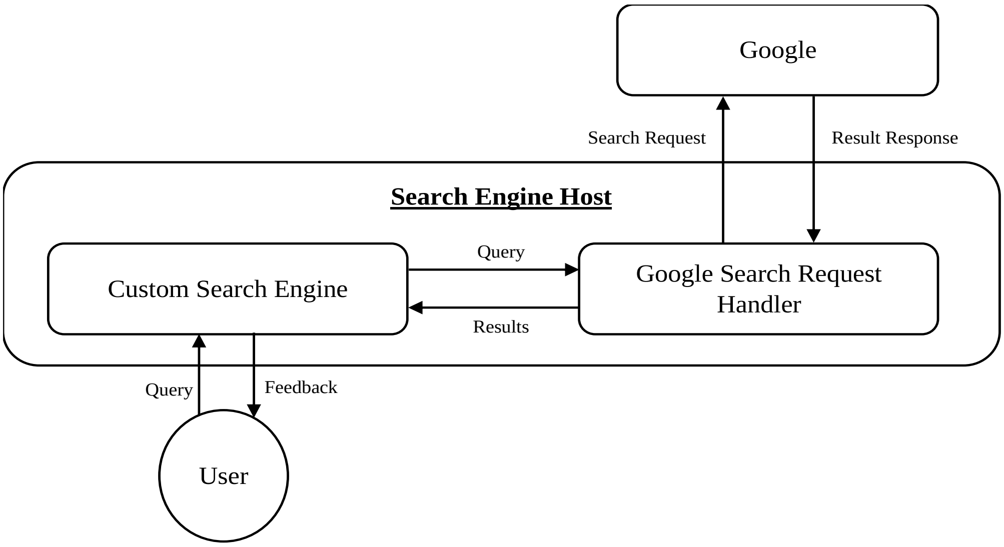


Fig.4. Custom Search Engine Architecture.

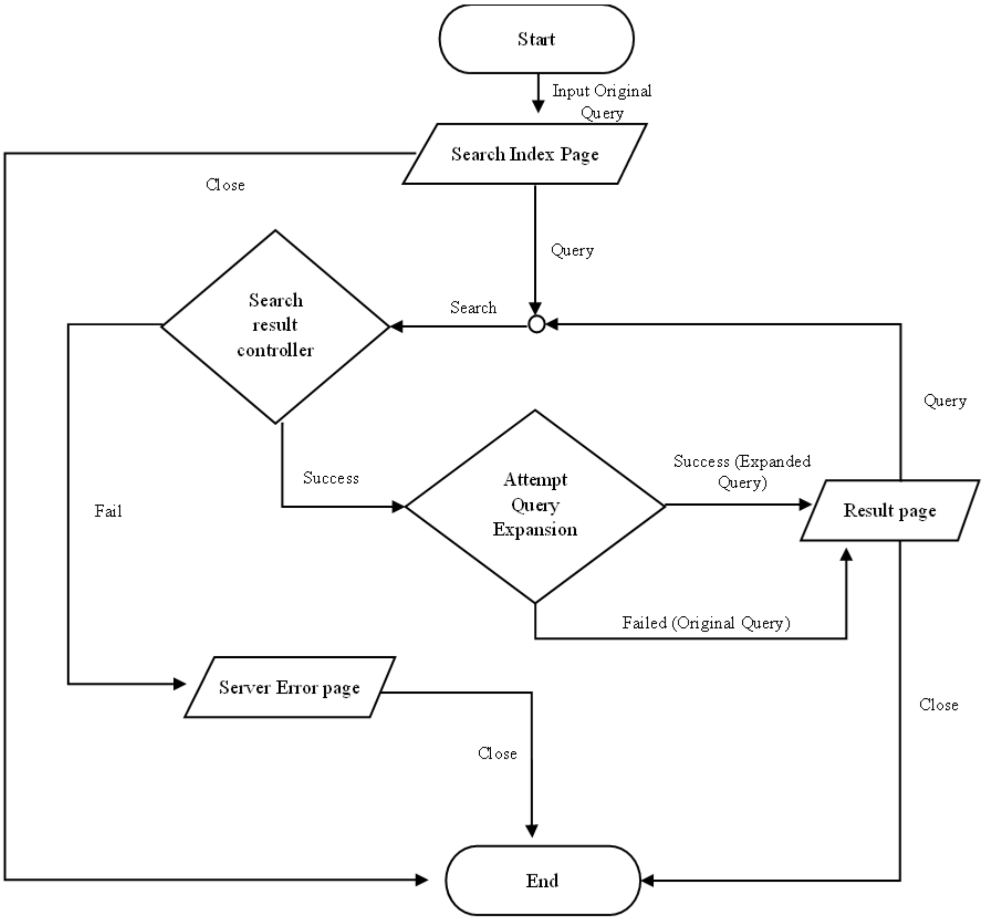


Fig.5. Query Expansion Search Engine Flowchart.

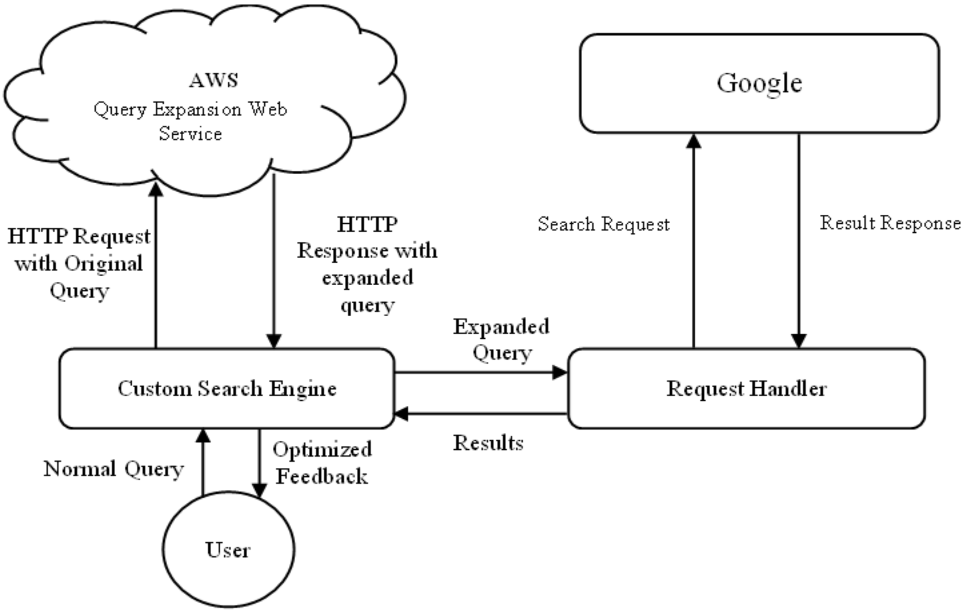
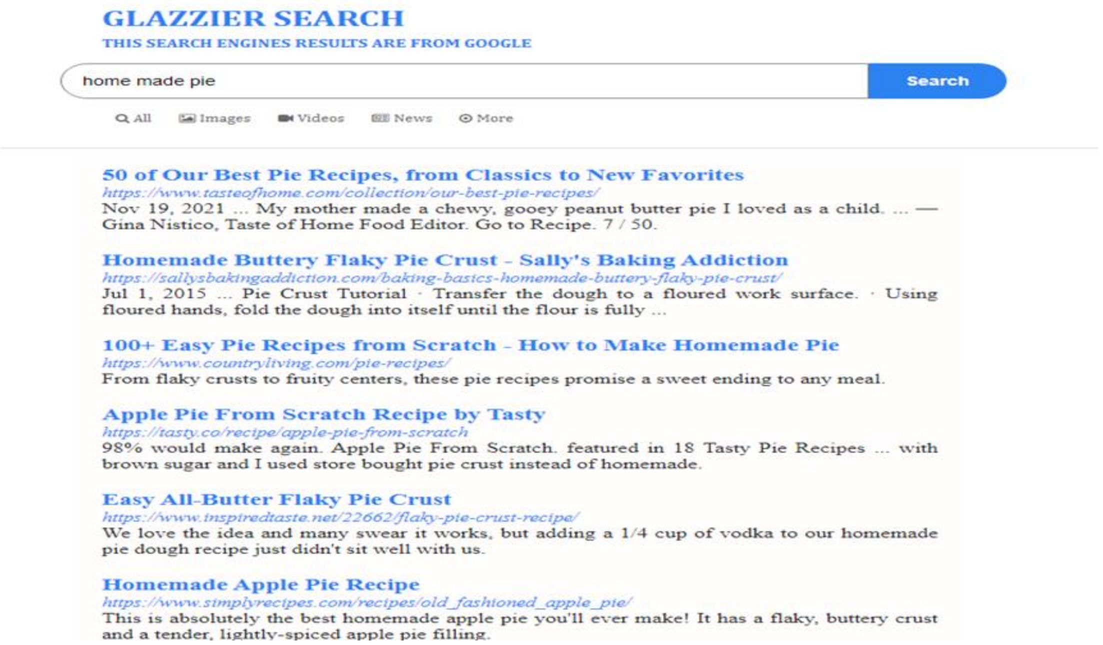


Fig.6. Search Engine with Query Expansion Architecture.

5. Evaluation

In this section, the evaluation of the proposed system is presented. We tested the *query expansion module* with about 50 queries and the expansion result were obtained. These queries are standard TREC queries 126 – 175 obtained from TREC’s website [34]. The queries in their original document each contain the following specifications, language, query number, title, description and narrative. We extracted the queries from the original document into a query file with each query on a row. Appendix A shows the table containing the results of our query expansion test. During the testing, our system was able to expand 50 queries in 10 minutes. This duration is relatively small considering the number of documents processed during expansion and performance optimization. The search result page for the custom search engine (without expansion feature) is shown in Fig. 7.



GLAZZIER SEARCH

THIS SEARCH ENGINES RESULTS ARE FROM GOOGLE

50 of Our Best Pie Recipes, from Classics to New Favorites
https://www.tasteofhome.com/collection/our-best-pie-recipes/
Nov 19, 2021 ... My mother made a chewy, gooey peanut butter pie I loved as a child. ... — Gina Nistico, Taste of Home Food Editor. Go to Recipe. 7 / 50.

Homemade Buttery Flaky Pie Crust - Sally's Baking Addiction
https://sallysbakingaddiction.com/baking-basics-homemade-buttery-flaky-pie-crust/
Jul 1, 2015 ... Pie Crust Tutorial · Transfer the dough to a floured work surface. · Using floured hands, fold the dough into itself until the flour is fully ...

100+ Easy Pie Recipes from Scratch - How to Make Homemade Pie
https://www.countryliving.com/pie-recipes/
From flaky crusts to fruity centers, these pie recipes promise a sweet ending to any meal.

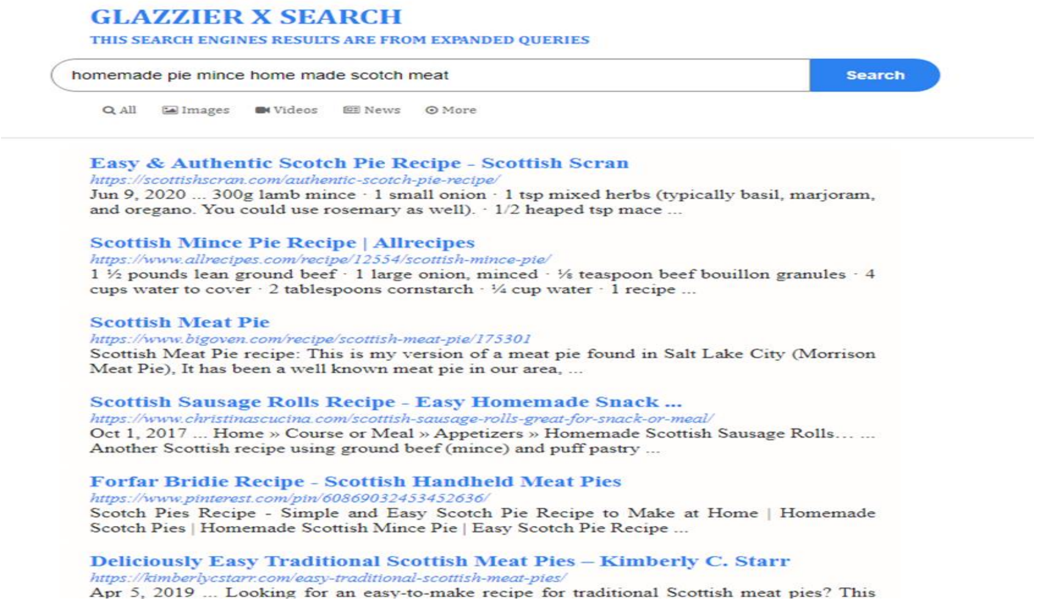
Apple Pie From Scratch Recipe by Tasty
https://tasty.co/recipe/apple-pie-from-scratch
98% would make again. Apple Pie From Scratch. featured in 18 Tasty Pie Recipes ... with brown sugar and I used store bought pie crust instead of homemade.

Easy All-Butter Flaky Pie Crust
https://www.inspiredtaste.net/22662/flaky-pie-crust-recipe/
We love the idea and many swear it works, but adding a 1/4 cup of vodka to our homemade pie dough recipe just didn't sit well with us.

Homemade Apple Pie Recipe
https://www.simplyrecipes.com/recipes/old_fashioned_apple_pie/
This is absolutely the best homemade apple pie you'll ever make! It has a flaky, buttery crust and a tender, lightly-spiced apple pie filling.

Fig.7. Search Engine Feedback without Expansion.

The search result page for the search engine with query expansion capabilities is shown in Fig. 8.



GLAZZIER X SEARCH

THIS SEARCH ENGINES RESULTS ARE FROM EXPANDED QUERIES

homemade pie mince home made scotch meat

Easy & Authentic Scotch Pie Recipe - Scottish Scran
https://scottishscran.com/authentic-scotch-pie-recipe/
Jun 9, 2020 ... 300g lamb mince · 1 small onion · 1 tsp mixed herbs (typically basil, marjoram, and oregano. You could use rosemary as well). · 1/2 heaped tsp mace ...

Scottish Mince Pie Recipe | Allrecipes
https://www.allrecipes.com/recipe/12554/scottish-mince-pie/
1 ½ pounds lean ground beef · 1 large onion, minced · ⅛ teaspoon beef bouillon granules · 4 cups water to cover · 2 tablespoons cornstarch · ¼ cup water · 1 recipe ...

Scottish Meat Pie
https://www.bigoven.com/recipe/scottish-meat-pie/175301
Scottish Meat Pie recipe: This is my version of a meat pie found in Salt Lake City (Morrison Meat Pie), It has been a well known meat pie in our area, ...

Scottish Sausage Rolls Recipe - Easy Homemade Snack ...
https://www.christinascucina.com/scottish-sausage-rolls-great-for-snack-or-meal/
Oct 1, 2017 ... Home » Course or Meal » Appetizers » Homemade Scottish Sausage Rolls... ... Another Scottish recipe using ground beef (mince) and puff pastry ...

Forfar Bridie Recipe - Scottish Handheld Meat Pies
https://www.pinterest.com/pin/60869032453452636/
Scotch Pies Recipe - Simple and Easy Scotch Pie Recipe to Make at Home | Homemade Scotch Pies | Homemade Scottish Mince Pie | Easy Scotch Pie Recipe ...

Deliciously Easy Traditional Scottish Meat Pies – Kimberly C. Starr
https://kimberlycstarr.com/easy-traditional-scottish-meat-pies/
Apr 5, 2019 ... Looking for an easy-to-make recipe for traditional Scottish meat pies? This

Fig.8. Feedback of Search Engine with Expansion.

We also conducted a comparative evaluation of the two systems: the *controlled system* without expansion feature, and the *experimental system* with the expansion future. Queries were entered in both controlled and experimental systems and the texts of each of the first 10 documents of both systems were copied and saved. To determine the quality of the documents retrieved, we used Cosine similarity measurement to get the similarity of the retrieved documents to the TREC relevance description of the query. The Cosine similarity weight then becomes the similarity weight or relevancy of that document to the query. The effectiveness of the method therefore greatly lies in the quality of the relevance description. The step-by-step evaluation procedure is as follows:

- Enter a query from the TREC dataset into the controlled and experimental systems
- Extract the texts of the first 10 documents retrieved from both systems and save them
- Load document texts from the controlled system
- Load document text from the experimental system
- Load the relevance descriptions for each query from the TREC dataset
- For each query *q*, get documents *D* from the controlled system and compare it with the relevance description for that query using cosine similarity
- Repeat the same process for the documents from the experimental system and make the comparison with the relevance description for the query using cosine similarity
- At the end of step 7, the cosine similarity measure (score) for each document from the controlled system for each query is obtained and saved, as well as the score from the experimental system.

The results of the evaluation for the controlled system (system without query expansion) and experimental system (system with query expansion) for one of the queries, “Price hike of petroleum products” is presented in Table 1. and depicted in Fig. 8.

Table 1. Similarity Comparison of the Controlled System and Experimental System.

Doc	Controlled System score	Experimental System score
doc1	0.13	0.16
doc2	0.09	0.28
doc3	0.12	0.05
doc4	0.1	0.06
doc5	0.13	0.06
doc6	0.26	0.14
doc7	0.1	0.16
doc8	0.02	0.13
doc9	0.09	0.05
doc10	0.05	0.11
Total	1.09	1.2

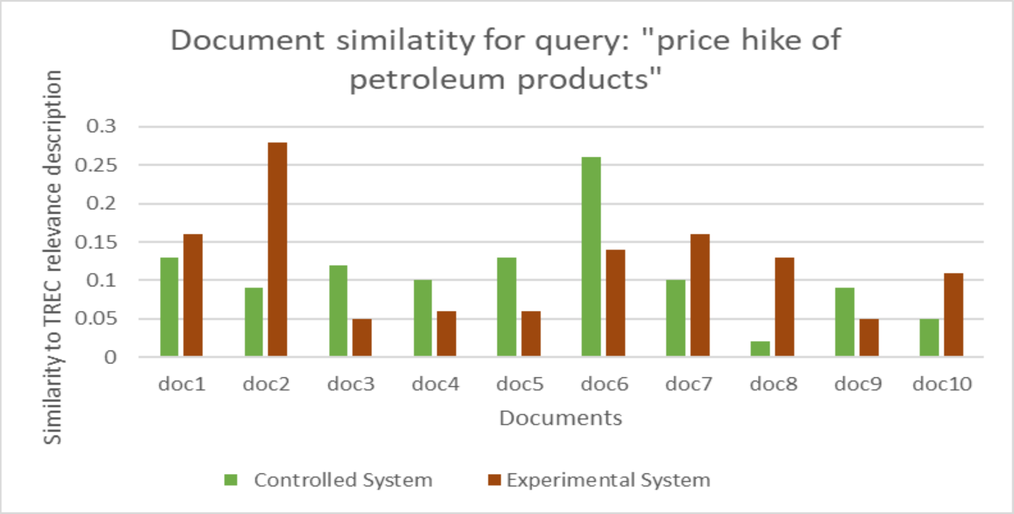


Fig.9. Similarity Comparison of the Controlled System and Experimental System.

The result shows that the higher the similarity value, the higher the document's relevance to the relevance description, implying that the document contains more useful information for the query. The evaluation as shown in Table 1 indicate that the experimental system with an expansion feature returned more relevant documents with an improvement of 11% as compared to the controlled system without the expansion feature.

6. Conclusion

In this research work, we developed a query expansion model based on Azak and Deepak's [18] model with modifications in Candidate Expansion Terms selection. In designing the system with query expansion, live Wikipedia data containing in-links and out-links articles were used to extract the best terms to append to the original query. Unlike the border rank approach used by [18] to retrieve the best Wikipedia documents, we chose Google Wikipedia search over border rank, and we also modified the model in terms of Wikipedia articles retrieval and document weighting using cosine similarity. Our expansion model returned results in a suitable time. We tested our expansion model on the TREC dataset (queries 126 – 175) and obtained well-expanded queries that returned relevant results. The comparative evaluation of the system shows an 11% improvement in the total results returned by the system with the expansion feature.

This work is limited in the area of language use. The only language considered in this research work is English. Future work will include an A/B comparative evaluation of our system and that of [18] using the FIRE dataset, and evaluation matrices like precision, recall and Mean Average Precision (MAP) will be employed for the evaluation.

Appendix A TREC 126 – 175 Expansion Results

S/N	Original Query	Expanded Query
1	Swine flu vaccine	Swine flu vaccine 2009 pandemic h1n1/09 virus united
2	Rare cosmic events	Rare cosmic events list future astronomical solar eclipse
3	Godhra train attack	Godhra train attack 2002 gujarat riots burning nanavati-mehta
4	Michael Jacksons untimely death	Michael jackson's untimely death (album) bad katherine
5	Price hike of petroleum products	Price hike of petroleum products oil gasoline diesel usage pricing
6	Abduction and murder of journalists	Abduction and murder of journalists daniel pearl james foley
7	Barack Obamas victory	Barack obama's victory electoral history joe biden george
8	Torture at the Abu Ghraib prison	Torture at the abu ghraib prison prisoner abuse iraq scandals rationale
9	Ban slapped on SIMI	Ban slapped on simi popular front india assault t.
10	Indias agriculture-friendly central budget	India's agriculture-friendly central budget 2020 union economy west bengal
11	Piracy in the world of entertainment	Piracy in the world of entertainment alliance creativity pirates magic kingdom
12	The death of LTTE head	The death of ltte head list commanders’ assassinations sri lankan
13	Indias Womens Reservation Bill	Indias womens reservation bill women 's suffrage india herabai
14	Vanquishing the Somali pirates	Vanquishing the somali pirates’ slavery ethiopia encomienda menelik ii

15	Search for life and water in space	Search for life and water in space astrobiology extraterrestrial mars panspermia abiogenesis
16	Birth of cloned human babies	Birth of cloned human babies brigitte boisselier clonaid raëlism raëlian
17	Illegal felling of trees	Illegal felling of trees logging madagascar deforestation 2009 malagasy
18	TATAs car, Nano	Tata car, nano motors cars group list entities
19	Bhopal gas tragedy	Bhopal gas tragedy disaster warren anderson (american
20	Assassination of Benazir Bhutto	Assassination of benazir bhutto zulfikar ali 2007 karsaz bombing
21	Ram Janmabhoomi verdict	Ram janmabhoomi verdict 2019 supreme court ayodhya dispute
22	Cybercrime in India	Cybercrime in india cybercrime convention computer security cyberwarfare
23	Popularity of social networking sites	Popularity of social networking sites media use politics service problematic
24	Commonwealth Games in Delhi	Commonwealth games in delhi 2010 concerns controversies venues 2014
25	Bill Gates philanthropic endeavours	Bill gates philanthropic endeavours melinda french sr . Microsoft
26	An Indian win the Nobel Prize for Chemistry	An indian wins the nobel prize for chemistry controversies physics physiology medicine peace
27	Successful missile test in India	Successful missile test in india anti-ballistic mission shakti anti-satellite weapon
28	Structure of the Solar System	Structure of the solar system universe formation evolution sun big
29	A. Raja and the 2G Spectrum scam	A. Raja and the 2g spectrum scam case list scandals india supreme
30	Attack on the Taj in Mumbai	Attack on the taj in mumbai 2008 attacks attribution reactions lashkar-e-taiba
31	Shashi Tharoor in the IPL controversy	Shashi tharoor in the ipl controversy sunanda pushkar 's oxford union
32	Successful Indian films	Successful indian films list highest-grossing bollywood india cinema
33	Americas attack on Afghanistan	America's attack on afghanistan war (2001–2021) history
34	Sheikh Hasina in the 2008 elections	Sheikh hasina in the 2008 elections bangladesh bangladeshi general election history
35	Iraq War 2003	Iraq war 2003 invasion rationale iraqi insurgency (
36	George Bushs anti-terrorism operations	George bush's anti-terrorism operations war terror operation enduring freedom
37	Catastrophic tornadoes	Catastrophic tornadoes 1953 waco tornado outbreak records
38	Tiger conservation in India	Tiger conservation in india reserves project bengal national authority
39	Michael Jackson and child abuse	Michael jackson and child abuse leaving neverland trial cultural impact
40	Marriage or divorce laws	Marriage or divorce laws law country no-fault united states
41	The Kanishka air disaster	The kanishka air disaster india flight 182 babbar khalsa
42	Chechen rebels and the government	Chechen rebels and the government republic ichkeria chechnya second war
43	Suu Kyi under house arrest	Suu kyi under house arrest aung san 54 university avenue
44	Naxalite attacks	Naxalite attacks timeline naxalite–maoist insurgency communist party
45	Development of Indian Hockey	Development of indian hockey field india league sardara singh
46	Prime witness in the Best Bakery case	Prime witness in the best bakery case 2002 gujarat riots godhra train
47	Female foeticide in India	Female foeticide in india infanticide sex-selective abortion gender inequality
48	2008 Olympics	2008 olympics summer concerns controversies olympic games
49	International economic slump	International economic slump great recession united states subprime
50	Sachin Tendulkars record of runs in Test Cricket	Sachin tendulkars record of runs in test cricket tendulkar list international centuries mohammad

References

[1] D. Di Caprio, F.J. Santos-Arteaga and M. Tavana, "An information retrieval benchmarking model of satisficing and impatient users’ behavior in online search environments," *Expert Syst.Appl.*, vol. 191, 1 April 2022, pp. 116352.

[2] A. Undu and S. Akuma, "Investigating the Usability of a University Website from the Users’ Perspective: An Empirical Study of Benue State University Website," *International Journal of Computer and Information Engineering*, vol. 12(10), pp. 922-929.

[3] M. Bouchakwa, Y. Ayadi and I. Amous, "An ambiguous tag-based query reformulation technique for an effective semantic-based social image research," *Procedia Computer Science*, vol. 176, 2020, pp. 508-520.

[4] J. Chen, J. Mao, Y. Liu, F. Zhang, M. Zhang and S. Ma, "Towards a Better Understanding of Query Reformulation Behavior in Web Search,", pp. 743-755.

[5] J. Mao, Y. Liu, K. Zhou, J. Nie, M. Zhang, S. Ma, J. Sun and H. Luo, "When does Relevance Mean Usefulness and User Satisfaction in Web Search?" *SIGIR '16 Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, July 17 -21, 2016, pp. 463-472.

[6] P. Ankalkoti, "Survey on Search Engine Optimization Tools & Techniques," *Imperial journal of interdisciplinary research,*

vol.3.

[7] S. Akuma, C. Jayne, R. Iqbal and F. Doctor, "Implicit predictive indicators: Mouse activity and dwell time," *IFIP Advances in Information and Communication Technology*, vol. 436, pp. 162-171.

[8] U. Kruschwitz, D. Lungley, M.-. Albakour and D. Song, "Deriving query suggestions for site search," *J Am Soc Inf Sci Tec*, vol. 64, pp. 1975-1994.

[9] M. Sanderson and W.B. Croft, "The History of Information Retrieval Research," *Proceedings of the IEEE, 100(Special Centennial Issue)*, pp. 1444-1451.

[10] M. Nagpal and J.A. Petersen, "Keyword Selection Strategies in Search Engine Optimization: How Relevant is Relevance?" *J.Retail.*, vol. 97, no. 4, December 2021, pp. 746-763.

[11] Stephen Akuma, Rahat Iqbal, "Development of Relevance Feedback System using Regression Predictive Model and TF-IDF Algorithm", International Journal of Education and Management Engineering(IJEME), Vol.8, No.4, pp.31-49, 2018.DOI:10.5815/ijeme.2018.04.04

[12] T. Kucukyilmaz, "Exploiting temporal changes in query submission behavior for improving the search engine result cache performance," *Information Processing & Management*, vol. 58, no. 3, May 2021, pp. 102533.

[13] D.K. Sharma, R. Pamula and D.S. Chauhan, "A Comparative Analysis of Fuzzy Logic Based Query Expansion Approaches for Document Retrieval," *In M. Singh, P. K. Gupta, V. Tyagi, J. Flusser, & T. Ören (Eds.), Advances in Computing and Data Sciences*, vol. 906, pp. 336-345.

[14] H.K. Azad and A. Deepak, "A novel model for query expansion using pseu-do-relevant web knowledge," *Inf.Sci.*, August 2019.

[15] C. Carpineto and G. Romano, "'A Survey of Automatic Query Expansion in Information Retrieval," *Information Retrieval. ACM Comput. Surv.*, vol. 44(1).

[16] J. Ooi, M. Xiuqin, Q. Hongwu and S.C. Liew, "A survey of query expansion, query suggestion and query refinement techniques,", pp. 112-117.

[17] N.J. Belkin, D. Kelly, G. Kim, J.Y. Kim, H.J. Lee, G. Muresan, M.C. Tang, X.J. Yuan and C. Cool, "Query Length in Interactive Information Retrieval,", pp. 205-2012.

[18] H.K. Azad and A. Deepak, "A new approach for query expansion using Wikipedia and WordNet," *Inf.Sci.*, vol. 492, August 2019, pp. 147-163.

[19] D. Pal, M. Mitra and K. Datta, "Improving query expansion using WordNet," *J. Assoc. Inf. Sci. Technol*, vol. 65(12), pp. 2469-2478.

[20] D. Roy, D. Paul, M. Mitra and U. Garain, "Using Word Embeddings for Automatic Query Expansion," *ArXiv:1606.07608 [Cs]*.

[21] A. Keikha, F. Ensan and E. Bagheri, "'Query expansion using pseudo relevance feedback on wikipedia," *Journal of Intelligent Information Systems*, vol. 50(1), pp. 1-24.

[22] R.K. Bisht and I.P. Bisht, "Effect of Query Formation on Web Search Engine Results," *International Journal on Natural Language Computing*, vol. 2(1), pp. 31-36.

[23] S. Akuma and R. Iqbal, "Investigation of Students' Information Seeking Behaviour," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 12, pp. 28-35.

[24] C. Claudio, D. Renato, R. Giovanni and B. Brigitte, "An Information Theoretic Approach to Automatic Query Expansion," *ACM Transactions on Information Systems*, vol. 19, pp. 1-17.

[25] D.K. Sharma, R. Pamula and D.S. Chauhan, "Query expansion – Hybrid framework using fuzzy logic and PRF," *Measurement*, vol. 198, July 2022, pp. 111300.

[26] D.B. Jake, H. HIlary and S. Maria, "User Preference and Search Engine Latency,".

[27] C. Xiong and J. Callan, "Query Expansion with Freebas,", pp. 111-120.

[28] T. Russell-Rose, P. Gooch and U. Kruschwitz, "Interactive query expansion for professional search applications," *Business Information Review*, vol. 38(3), pp. 127-137.

[29] Q. Yonggang and H. Frei, "Concept based query expansion,", pp. 160-169.

[30] C. Hang, W. Ji-Rong, N. Jian-Yun and M. Wei-Ying, "Probabilistic query expansion using query logs,", pp. 325-332.

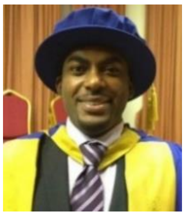
[31] B.M. Fonseca, P. Golgher, B. Pôssas, B. Ribeiro-Neto and N. Ziviani, "Concept-based interactive query expansion," *Concept-based interactive query expansion*, pp. 696.

[32] S. Riezler, Y. Liu and A. Vasserman, "Translating queries into snippets for improved query expansion," *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08*, pp. 737-744.

[33] A. Keikha, F. Ensan and E. Bagheri, "Query expansion using pseudo relevance feedback on wikipedia," *Journal of Intelligent Information Systems*, vol. 50(1), pp. 1-24.

[34] TREC, "TREC Queries 126 – 175,", vol. 2022.

Authors' Profiles



Stephen Akuma is a Lecturer in the Department of Mathematics, Computer Science and Statistics and Deputy Director at the Center for Open and Distance Learning at Benue State University. He is an experienced research scientist and a university Lecturer with a demonstrated history of working in data science, human-computer interaction and information retrieval, focusing on user modelling and personalization. He holds a PhD in Computing and a Master's degree in Software Development (Distinction) from Coventry University, United Kingdom. Stephen also obtained a Bachelor's degree in Computer Science (2.1) from Benue State University. His research area is informational retrieval, personalization and machine learning. He has over 13 years of Teaching Experience and has published papers in reputable International Conferences and Journals.

Promise Anendah completed his undergraduate programme at the Department of Mathematics, Computer Science and Statistics, Benue State University with a BSc (Hons) degree in Computer Science (2.1). He is presently working as a Software Developer at SACS Computers (www.sacscomputers.com), Makurdi, Nigeria.

How to cite this paper: Stephen Akuma, Promise Anendah, "A New Query Expansion Approach for Improving Web Search Ranking", International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.1, pp.42-55, 2023. DOI:10.5815/ijitcs.2023.01.05