Stanchev, et al., 2019. Published in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 150-153.

Combining Parts of Speech, Term Proximity, and Query Expansion for Document Retrieval

Eric LaBouve* Lubomir Stanchev†
Department of Computer Science and Software Engineering
California Polytechnic State University
San Luis Obispo, CA, USA
*elabouve@calpoly.edu †lstanche@calpoly.edu

***Abstract*—Document retrieval systems recover documents from a database and order them according to their perceived relevance to a user's search query. This is a difficult task for machines to accomplish because there exists a *semantic gap* between the meaning of the terms in a user's literal query and a user's true intentions. The main goal of this study is to modify the Okapi BM25 document retrieval system to improve search results for textual queries and unstructured, textual corpora. This research hypothesizes that Okapi BM25 is not taking full advantage of the structure of text inside documents. This structure holds valuable semantic information that can be used to increase the model's accuracy. Modifications that account for a term's part of speech, the proximity between a pair of related terms, the proximity of a term with respect to its location in a document, and query expansion are used to augment Okapi BM25. The study resulted in 87 modifications which were all validated using open source corpora. The top scoring modification from the validation set was then tested under the Lisa corpus and the model performed 10.25% better than Okapi BM25 when evaluated under mean average precision.**

Keywords: Semantic Analysis, Document Retrieval, Query Expansion, Term Proximity, Search, Okapi BM25

I. Introduction

One of the most pervasive document retrieval engines in everyday society is Google's search engine. Google's search engine works well because documents on the Internet are highly structured with HTML elements and RDF triples that explicitly define the contents of web pages. These metadata elements are used as training data by Google's page rank algorithm [10] to score a document based on the document's popularity among other web pages. Due to the technical nature of the page rank algorithm, Google's search engine fails when users are trying to search for documents that are unpopular. This is a problem because documents can be both unpopular and relevant to the user's search query. As a result, popular documents are circularly discovered by many individuals and unpopular, yet relevant, documents go unnoticed. The solution is to develop a system that scores a document based on its content rather than its perceived popularity among other documents in the same corpus.

It is important to research alternative ways to rank documents for a handful of reasons. First, systems that primarily rely on training data will not operate well if the domain of the training data is disjoint from the domain of the deployment environment [14]. Second, a search engine that is agnostic to any preexisting document structure, such as HTML elements, RDF triples, or bibliographic citations [16], could be applied to a larger set of corpora. The Okapi BM25 document retrieval system [13] does not require any training data and it does not depend on any preexisting document structure.

Ranking documents is a difficult problem because the context of a query is only partially observable. For example, there commonly exists a mismatch, known as the *semantic gap* [17], between the user's literal query and the true meaning behind what the user intended to type. Additionally, understanding the true nature behind a user's query is made more difficult because languages are dynamic with respect to time and culture. For these reasons, developing the perfect document retrieval system is much like designing a black box where the true relevance rating for a document is not observable. To minimize this uncertainty, we use publicly available information retrieval collections with relevance ratings that have been determined by human evaluation.

Previously proposed upgrades to Okapi BM25 are inadequate because they only optimize the model on a small number of parameters and their experiments rely on data sets that do not have standardized relevance ratings. Little research has been done to optimize Okapi BM25 across many modification themes. This paper hypothesizes that Okapi BM25 can be modified to take advantage of many contextual themes, such as a term's part of speech, the proximity of related terms to each other, the proximity of terms within a document, and query expansion techniques to improve the model's accuracy. Our paper's unique contribution is a version of the Okapi BM25 system that takes advantage of a wide variety of contextual information inside text documents. The system was built by designing and validating a large number of modifications against four corpora: Cranfield, Adi, Medline, and Time, and then testing the best modification against the Lisa corpus. The results show that the new model positively increases the mean average precision (MAP) of the original Okapi BM25 model by 10.25% and takes advantage of parts of speech, term to term proximity, term to document proximity, and query expansion.

II. Related Research

A noteworthy attempt to expand Okapi BM25 was conducted by Cummins *et al.* [4]. The researchers used a genetic algorithm to evolve the model to favor high MAP scores

when trained against 69,500 documents and 55 queries. Their resulting model relied on distance proximity measures between pairs of terms. Other research has shown that emphasizing the distances between pairs of grammatically related terms, such as compound nouns, can result in higher precision values and that the appearance of a single term holds little semantic meaning unless it is found near its related term(s) [2].

Some researchers have found success when analyzing spans, which are segments of text from a document that incorporates all query terms, or a subset of the query terms. Successful experiments have focused on the first occurrence of query terms in a document [12] and designing sophisticated term frequency measurements that focus on the density of nonoverlapping spans [15]. Other successful modifications take advantage of a term's position in a document, such as the research done by Blanco *et al.*, to generalize BM25F [11] to unstructured text [3]. Their approach splits a document into "virtual regions", much like a spans, and weights the terms in these regions proportionally to the section's statistical significance.

The last significant theme of modifications is query expansion, which is an attempt to add related terms to a query in order to express the original query in a more detailed way. There are three major areas of query expansion as identified by Ooi *et al.* [9]: query expansion using corpus dependent knowledge models, query expansion using relevance feedback, and query expansion using language models. Some researchers have found success with query expansion [5], while other researchers have concluded that query expansion will inevitably hurt a system's recall due to vocabulary mismatch or a system's precision due to topic drift [1].

The related research that is presented shows that Okapi BM25 can be improved when optimized for a single modification theme, but there is limited research on ways to optimize against multiple modification themes. The remainder of this paper will demonstrate how a variety of modification themes can be combined to improve Okapi BM25.

III. Solution / Implementation

This section details how Okapi BM25 is extended to enable many modifications and describes four modification themes that were tested. The first theme analyzes a query term's part of speech. The second theme analyzes the distance between pairs of query terms. The third theme analyzes the position of a single query term with respect to its location within a document. Finally, the fourth theme explores methods for query expansion.

A. Extending Okapi BM25

A more extensible version of Okapi BM25 can be built to utilize many modifications. The score generated for a single term is modified to include a collection of boosts, which are proportional to the absolute value of the term's original Okapi BM25 score. Each activated modification contributes a single boost value and these boost values are added to the term's original Okapi BM25 score. Equation 1 is the boosting function used for all modifications. In Equation 1, *OkapiBM25* is the original score calculated from Okapi BM25 and *Influence* is a modification specific value that is responsible for scaling a term's score. *Influence* values range from zero to two and are either determined through training, chosen heuristically, or computed algorithmically.

$$Boost = (Influence - 1) \cdot |OkapiBM25| \tag{1}$$

B. Parts of Speech

The simplest set of modifications is to scale up or down the *Influence* of an individual term according to its part of speech. In order to simplify contextual analysis, words are assumed to only be nouns, verbs, adjectives, or adverbs. *Influence* values for each part of speech take on values that are both greater than one and less than one. The values are set after training the modifications on the Cranfield corpus until a local maximum MAP value is reached.

C. Term to Term

We take inspiration from [2] and assume that pairs of query terms are related when an adjective or adverb is found next to a noun or verb. The idea behind this assumption is that a modifying term contains the most semantic meaning if found near its corresponding subject. For example, in the query "Red cars for sale", the term "Red" is semantically insignificant if it is found in a document that does not contain the word "car."

Three different sets of modifications are built to evaluate pairs of semantically related terms. The first set excludes the score from modifiers unless the term that immediately follows in the document is the corresponding subject. The second set rewards a document for containing bigrams constructed from the query. Bigrams are assembled using one of three different techniques: between adjacent terms, between adjacent adjectives and nouns, or between adjacent adverbs and verbs. The *Influence* value for each technique is determined by training sample queries on the Cranfield corpus until local maximum MAP values are reached. The third set is designed to boost nonadjacent modifiers and subjects. Equation 2 is used to determine the *Influence* value between the two nonadjacent query terms, where x is the minimum distance between a pair of query terms in a document calculated as the difference between their absolute indexes.

$$Influence = max(-\frac{x}{4} + 2, 1) \tag{2}$$

D. Term to Document

For the next set of modifications, we propose that if users expect relevant information to appear at the start of documents, then a document should be rewarded for containing query terms closer to the front of the document. Equation 3 is used to reward terms based on a term's first occurrence in a document.

$$Influence = \frac{2 * dl_j - idx_i}{dl_j} \tag{3}$$

Equation 3 is a linear function, where dl_j is the length of document j, measured as the sum of all its terms and idx_i is the absolute index location of term i, where the first term in

the document has an idx_i value of zero. The upper bound for the function is heuristically set to two because terms at the front of a document are assumed to be twice as important as terms that appear at the end of a document.

E. Query Expansion

Three methods for global query expansion are implemented. For each query expansion method, a query term will be awarded one boost value for each expansion term. Unlike previous modifications, query expansion boost values are computed as the expansion term's original Okapi BM25 score multiplied by the specified *Influence* value.

The first method uses the APIs exposed by WordNet [8]. Using the APIs is nontrivial because words may have multiple definitions and parts of speech. In order to look up the correct word in WordNet and extract cognitive synonyms, the Lesk algorithm [6] is used to perform word sense disambiguation. Unfortunately, WordNet does not provide the strength of the similarity between a term and its cognitive synonyms. So, we set the *Influence* value for all WordNet API expansion terms to 0.9 because expansion terms will have a slightly lower probability of being relevant than the original query term.

Although WordNet does not quantify the similarity between terms, recent research shows that similarity scores can be derived if the WordNet database is arranged in a probability graph [17]. Semantically similar terms are discovered from the probability graph by computing random walks from the node that represents the unexpanded query term. After computing many random walks, the nodes that are traversed most often represent the semantically similar terms. The similarity score between the unexpanded term and an expansion term is then the proportion of times the expansion term's node was visited in the random walks. This proportion is then used as the expansion term's *Influence* value.

The last query expansion category uses word vectors generated using the Word2Vec algorithm [7] on the Google News corpus[1]. Since words are represented as vectors, the cosine similarity equation can be used to quantify the similarity between words. This similarity score is then used as the expansion term's *Influence* value.

IV. Experimental Procedure

Modifications are validated against four publicly available benchmarks[2]: Cranfield, Adi, Medline, and Time. Each benchmark contains a set of documents, a set of queries, and an exhaustive list of relevance scores for all query-document pairs. In total, there are just under 3,000 documents and 373 queries in the validation set. The Lisa benchmark is used as the testing set and contains 5,872 documents and 35 queries.

The experimental procedure is split up into three validation rounds and a fourth testing round. Round one runs the modifications independently. Round two combines the modifications within the same theme. Round three combines the modifications across multiple themes. Once all three validation rounds are completed, a single modification is selected for testing. The best performing modification is the one resulting in the highest sum of differences between the modification's MAP scores and the unmodified Okapi BM25 system's MAP scores across each benchmark in the validation set B, as shown in Equation 4.

$$\sum_{b \in B} (MAP(mod, b) - MAP(Okapi\ BM25, b)) \quad (4)$$

V. Results

The best model from the part of speech themed modifications increased the *Influence* for nouns and adjectives. Generally, decreasing the *Influence* of adjectives and adverbs and increasing the *Influence* of nouns positively affected the model's precision.

Modifications that measured the distance between terms resulted in relatively small changes in MAP because the probability of two terms with specific parts of speech appearing chronologically near each other in a document is a rare event. All these modifications resulted in lower precision values, except for the modifications that were designed to boost nonadjacent modifiers and subjects.

The modifications that targeted the position of a term in a document had the most positive impact on the validation benchmarks. Generally, modifications that targeted parts of speech that compose a larger majority of a document resulted in larger swings in accuracy, and vice versa. The results show that either targeting all parts of speech or just nouns and adjectives positively affected the model's precision.

Query expansion themed modifications had minor effects on the model's MAP score because only a few expansion terms were discovered for each query. Even when all three expansion techniques were combined to increase the number of expansion terms discovered for each query, the modification still lead to a decrease in the model's MAP score. We hypothesize that this is most likely due to topic drift. Although changes in precision were small, the two best performing query expansion techniques were when the WordNet Graph was used to expand only the nouns or when the WordNet Graph was used to expand terms that scored the lowest inverse document frequencies.

All together, 87 models were created across all validation rounds. From this set, the top scoring model was determined using Equation 4 and it was discovered that the top scoring model was created in validation round three. This model combines three modification themes. From the term to term theme, the model rewards adjectives and nouns for occurring near each other. From the term to document theme, the model rewards nouns and adjectives for appearing closer to the start of a document. Lastly, from the query expansion theme, the model uses the WordNet Graph to expand the terms that scored the lowest inverse document frequencies.

This model and the unmodified version of Okapi BM25 were then tested using the Lisa benchmark. When Okapi BM25 was ran against Lisa, the resulting MAP value was 0.357 and when the top model was ran against Lisa, the

[1]https://github.com/mmihaltz/word2vec-GoogleNews-vectors

[2]http://ir.dcs.gla.ac.uk/resources/test_collections

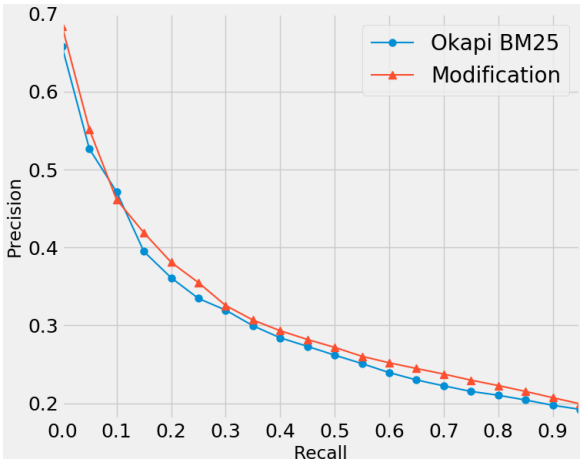

Fig. 1. Precision-recall curves for Okapi BM25 and the top model at recall bucket sizes 0.05.

top model scored a MAP value of 0.393. The difference between these results represents a 10.25% improvement. After inspecting the Lisa corpus for potential biases, document titles were removed from the benchmark to keep inherent document structure to a minimum. Both models were then reran against Lisa. This time, Okapi BM25 scored a MAP value of 0.304 and the top model scored a MAP value of 0.326, representing a 7.31% increase in performance.

The performances between the two models can also be compared using weighted average recall, where the recall scores are weighted proportionally to the number of relevant documents in each query. When ran against Lisa, Okapi BM25 returned recall scores of 0.145, 0.237, and 0.332 and the modified model returned recall scores of 0.155, 0.224, and 0.343 on the first 5, 10, and 20 documents returned for each query, respectively. From the first 5 documents returned, the top model obtained a recall that was 7.27% better than Okapi BM25. Then the recall dipped below Okapi BM25 once 10 documents were returned by around -5.56%. However in the long term, after 20 documents were returned, the top model returned a recall score that was 3.17% better than Okapi BM25.

The performances of both Okapi BM25 and the top model can be displayed on a precision-recall curve to gain more granular insight into how the MAP score is affected by the weighted average recall. In Figure 1, the blue line with dots represents Okapi BM25 and the red line with triangles represents the top model. From the graph, it is clear that the modified system scores a higher MAP value than Okapi BM25 at all recall levels, except at the recall range between 0.08 and 0.12. Despite this small range of values, the top model consistently outperforms the original Okapi BM25 model at short term and long term recall levels.

VI. Conclusion and Future Research

We have demonstrated a process to derive and validate many modifications for Okapi BM25. From the models that were created, the best performing model was selected and tested against the Lisa benchmark. This model combines query expansion, term to term proximity, and term to document proximity across various parts of speech. In conclusion, a model that combines many modification themes can be built to outperform Okapi BM25 in MAP and weighted average recall for most recall levels. One area for future research would be to extend Okapi BM25 to take advantage of more sophisticated natural language processing and grammar rules. For example, conjunction words can be identified to help locate the main subject of multi-clause sentences. The subject terms can then be weighted proportionally to their perceived significance.

References

[1] B. Al-Shboul and S. H. Myaeng. Analyzing topic drift in query expansion for information retrieval from a large-scale patent database. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 177–182, Jan 2014.

[2] M. P. S. Bhatia and A. Kumar. Contextual paradigm for ad hoc retrieval of user-centric web data. *IET Software*, 3(4):264–275, August 2009.

[3] R. Blanco and P. Boldi. Extending bm25 with multiple query operators. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 921–930, New York, NY, USA, 2012. ACM.

[4] R. Cummins and C. O'Riordan. Learning in a pairwise term-term proximity framework for information retrieval. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 251–258, New York, NY, USA, 2009. ACM.

[5] S. Kuzi, A. Shtok, and O. Kurland. Query expansion using word embeddings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1929–1932, New York, NY, USA, 2016. ACM.

[6] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26, New York, NY, USA, 1986. ACM.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[8] G. A. Miller. Wordnet: A lexical database for english, 1995.

[9] J. Ooi, X. Ma, H. Qin, and S. C. Liew. A survey of query expansion, query suggestion and query refinement techniques. In *2015 4th International Conference on Software Engineering and Computer Systems (ICSECS)*, pages 112–117, Aug 2015.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[11] J. R. Pérez-Agüera, J. Arroyo, J. Greenberg, J. P. Iglesias, and V. Fresno. Using bm25f for semantic search. In *Proceedings of the 3rd International Semantic Search Workshop*, SEMSEARCH '10, pages 2:1–2:8, New York, NY, USA, 2010. ACM.

[12] M. I. Rafique and M. Hassan. Utilizing distinct terms for proximity and phrases in the document for better information retrieval. In *2014 International Conference on Emerging Technologies (ICET)*, pages 100–105, Dec 2014.

[13] S. Robertson and S. Walker. Okapi/keenbow at trec8. In *The Eighth Text REtrieval Conference (TREC8)*, page 151162. Gaithersburg, MD: NIST, January 2000.

[14] H. Sanders and J. Saxe. Garbage in, garbage out: How purportedly great ml models can be screwed up by bad data. Technical report, July 2017.

[15] R. Song, J.-R. Wen, and W.-Y. Ma. Viewing term proximity from a different perspective. Technical report, May 2005.

[16] L. Soulier, L. Ben Jabeur, L. Tamine, and W. Bahsoun. Bibrank: A language-based model for co-ranking entities in bibliographic networks. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, pages 61–70, New York, NY, USA, 2012. ACM.

[17] L. Stanchev. Creating a similarity graph from wordnet. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, WIMS '14, pages 36:1–36:11, New York, NY, USA, 2014. ACM.