A Framework For Aggregating And Retrieving Relevant Information Using TF-IDF And Term Proximity In Support Of Maize Production

Philemon Kasyoka, Waweru Mwangi, Michael Kimwele

Abstract: This paper presents a framework for aggregating and retrieving relevant maize information using Term Frequency Inverse Document Frequency and Term Proximity. The framework aggregates information from agricultural websites and blogs through the use of RSS technology. Term Frequency Inverse Document Frequency is able to retrieve relevant documents from the aggregated RSS feeds however; the presence of a query term within a retrieved document does not necessarily imply relevance. Documents with same similarity score do not necessarily have the same level of relevance. To mitigate that problem we implement a term proximity scoring approach that will be able to improve relevance in the top-*k* documents returned by TF-IDF. The approach for term proximity score uses both the span-based method and pair-based method to ensure effective proximity scoring. User preference profile is based on keywords which form user query while text documents are composed of RSS description content and RSS title tag content. Stemming is applied on query and document terms for better precision. This framework will ensure maize farmers get the most relevant information from online sources.

Index Terms: Inverse Document Frequency, Information Retrieval, RSS, Term Frequency, Term Proximity

1.0 INTRODUCTION

Information Communication Technology has become a critical factor in driving growth and productivity in global economies. Farming is becoming a more time-critical and information-intense business. A push towards higher productivity will require an information-based system able to give the most relevant information effectively and with ease. According to [7] the cost of information from planting decision to selling at the wholesale market can make up to 11% of total production costs. With great advancement in technology and availability of affordability portable devices that can access internet, more farmers have become regular seekers of online information. Finding relevant information online is not an easy task and there is a need for an effective framework that can aggregate and retrieve relevant information. Most web developers are increasingly using the RSS technology to publish content on websites because of its capability of delivering up to-date posting. There is a need for a framework that will filter online agricultural information delivered through the use of RSS feeds to ensure farmers get the most relevant content based on their preferences.

With greater technological development in the area of Information Retrieval, Term Frequency-Inverse Document Frequency has been adopted as term weighting method for retrieval of relevant information in a vector space model. TF-IDF has a discriminatory power that allows a retrieval engine to quickly find relevant documents this makes it perfect for forming the foundation for more complex algorithms in information retrieval. During document retrieval users always expect that the most relevant documents to form the top-*k* list based on their preferences. The fact that a user query term appears in a text document it does not necessarily mean that the text document is relevant to their preferences, document with same similarity score need also to be ranked based on relevance to user preferences, it is for such reasons why in this paper term proximity is implemented on top documents returned by a retrieval model. Recent research by [10, 17] has shown that proximity between terms on a document is useful measure for improving retrieval quality in various information retrieval models. According to [4] relationship between terms plays an important role in text processing. Term proximity is introduced to TF-IDF to ensure only the most relevant documents are delivered to users. Through such as framework farmers are guaranteed to effectively get the most relevant maize information. The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 explains the proposed method of implementing the framework. Section 4 describes the test result and analysis. Section 5 covers conclusion and future work.

2.0 RELATED WORK

Research conducted by [11] analyzed web contents aggregated using RSS Technology, user profile was created from browsing user history and content was retrieved based on a comparison between user profiles characters of contents. They used Term Frequency (TF) to analyze web content. Its weakness was that the use of Term Frequency provides local weight of terms and not global weight, ignoring the global weight of a term would not retrieve the most relevant text documents. A similar framework was developed by [6] that recommended RSS web content using weighted TF-IDF. He concentrated on the construction of RSS

- *Philemon Kasyoka School of Computing and Information Technology Jomo Kenyatta University of Agriculture and Technology, P. O. BOX 62000-00200 Nairobi, Kenya, Email: pkasyoka@gmail.com*
- *Professor Waweru Mwangi, School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, P. O. BOX 62000- 00200 Nairobi, Kenya Email: waweru_mwangi@icsit.jkuat.ac.ke*
- *Doctor Michael Kimwele, School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, P. O. BOX 62000-00200 Nairobi, Kenya, Email: kimwele@icsit.jkuat.ac.ke*

document considering the channel element and item element. Term weight of title was calculated using the average TF-IDF of the entire document where a user query term was found in the title. Research on adhoc retrieval was done by [18] where proximity function was incorporated in BM25 to improve the performance of short queries. A window of size of five or less was used. They show that marginal improvements on larger retrieval on TREC data can be achieved. A similar proximity function using bi-term was developed by [15] their function permitted the use of arbitrary window size extracted from each scored document. More recently some approaches have been successful in employing proximity into a number of keyword based retrieval functions [16], their research has shown that proximity is important since significant improvements have been achieved for short queries. A framework for incorporating information about the proximity between all query terms into a TF-IDF retrieval model was developed by [14], the approach used in this paper calculates the proximity scores on top documents returned by the retrieval model by integrating both the pair-based and span-based approach. An approach for span was theorized by [13] where they considered an ordered list of query terms on a document and their position, the approached used in this paper identifies all spans found within a document and pick only the minimum span and calculates the proximity score for document. Most of the research modifies the query and document by removing stop words, a good information retrieval system should be able to perform well with or without stop-word, the approached used in this paper does not remove the stop-words. To find out whether term proximity can improve retrieval effectiveness we will use TF-IDF as our baseline.

3.0 METHODS

3.1 RSS AGGREGATION PROCESS

The process for aggregating maize information makes use of RSS Technology where information in form of RSS feeds is pooled in from agricultural websites or blogs. RSS is written in XML programming language and it is structured in channel where each channel consists of items with title, link, description and pubDate. The aaproach used in this paper extracts the title, description, link and pubDate to form text documents.

3.2 TERM FREQUENCY INVERSE DOCUMENT FREQUENCY

TF-IDF is the most common weighting method used to describe documents in the Vector Space Model. It is composed of Term Frequency and Inverse Document Frequency. Term frequency is a local weight of a term on a document achieved by counting the number of times a particular term occurs in a text document. The higher the frequency of a term on a document the more relevant the term is in that particular document. Inverse Document Frequency (IDF) is the count of all documents in the corpus divided by the number of documents that contains at least a single occurrence of the query term. The IDF gives a global view of the term across the entire corpus, the lower the IDF value the more significant the term will be and it is calculated using (1).

$$idf = \log\left(\frac{N}{df_i}\right) \qquad (1)$$

To get an effective term weight score, Term Frequency is combined with IDF by simply multiplying the values as shown by (2).

$$w_{t,d} = (1 + \log(\mathrm{tf}_{t,d})) \text{ x } idf \qquad (2)$$

Where $w_{t,d}$ is the term weight, $tf_{t,d}$ is number of occurrences of a specific term in a document and *idf* is the inverse document frequency. A study conducted by [1,8] found cosine similarity performs better than other correlation measures. Cosine remains the most effective measure for the visualization of the vector space as it is defined in a geometrical perspective. Cosine correlation coefficient is used to calculate similarity between query vector and document vector.

3.4 TERM PROXIMITY

Recent research [13,15] has shown that retrieval effectiveness can be greatly improved by integrating term proximity scores into the retrieval model. Effectiveness and relevance are critical concerns for information seekers, they expect information retrieval systems to return only what they need and as quickly as possible. According to [9] term proximity score is inversely proportional to the square of their distance within the document. In this paper we will integrate term proximity score with TF-IDF to ensure the top-*k* results are ranked based on relevance to user query terms.

3.4.1 TERM PROXIMITY CALCULATION

The notion of term proximity is to measure distance between terms found in a document; retrieval effectiveness can be greatly improved by integrating term proximity score into a retrieval model [15]. When there is a small distances between query terms it implies a strong semantic associations and when distances between query terms on a document is large, the query terms can be said to have a loose association or no association. According to [16] there are two approaches to performing term proximity and they are span-based approach and pair-based approach. In a study done by [5] they defined span as the length of a document fragment that covers all query term occurrences. In the pair-based approach distance between individual query terms occurrences is calculated, aggregated and transformed to a term proximity score. In a study conducted by [14] they identified different pair-based approaches that have been used in term proximity, their research indicated that minimum distance is highly correlated with relevance and therefore performs better than other methods, this corresponds with previous research work by [16]. In a study conducted by [10] they found that use of span-based term proximity can lead better performance than other approaches of calculating proximity. In their research span-based proximity led to significant gain while used with ranking models. A proximity function that incorporates other proximity measures may outperform other proximity approaches [14]. In this research the approach used to calculate term proximity is a hybrid approach that will use both the span-based approach and pair-based approach to maximize on the strengths of each approach. Minimum proximity distance denoted as *min_dist* will be used for pair-based approach and minimum span denoted as *min_span* will be used for span-based approach. The span-based approach used here is closely related to the one used by [3] where he defines the concept of matching span. His definition is ambiguous, instead of defining the minimal matching span as the smallest segment that contains all query terms occurring in a document at least once, his span is checked to ensure that it does not contain another sub matching span within it. His approach does not effectively address the distribution of query term throughout the document. Research by [16] shows that span based approach works better when normalized by query terms. [3] does not normalize the minimal matching span with the number of query terms found within the matching span. Our approach to span is slightly different, to get a *min_span* we will begin by counting the number of query term existing on the document, then scan the document for query term starting from left to right, when all the query term have been found the segment of the document will form a span, we will begin checking for the next span from the next term until we find all the spans within the document. The next step will be to select the shortest span to become our *min_span* and normalize its distance with the number of unique query terms on the document. Example, given sample document *Doc*={A,J,D,C,P,T,D,X,Q,T} and Query={A,D,T} where alphabets stand for different terms on the document

Pos	1	2	3	4	5	6	7	8	9	10
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Doc	A	J	C	D	P	T	D	X	A	T

Fig.1. Term Position Mapping

Fig. 1 shows how document terms are mapped against positions on a document, term position are counted from left to right starting from position 1 and so on. The spans for query {A,D,T} are {1,4, 6}=5 and {7,9,10}=3 ,since the order of query term is not relevant there the *min_span* is formed by query terms in positions 7,9 and 10. When getting the distance in pair-based approach only the minimum distance between query terms will be considered, still using the example document and query for terms A, D will consider distance from position 7 and 9 and not position 1 and 4, for terms A,T we will look at position 9 and 10 and ignore position 1 and 6, for terms D,T we will consider position 4 and 6 and ignore positions 7 and 10. The pair-based term proximity in this research is implemented using term scoring function that is closely related to the one defined by [18] where a feature that is proportional to the inverse square of the distance between each pair of queried terms is used. They limited query term consideration to a sliding window of size 5 and they calculated distance between all query terms occurrences on the document. Using such an approach the number of extracted pair dependencies grows exponentially with the number of query terms making longer queries impractical. The scoring function used in this paper is differs from their work as the denominator in the term scoring function used to calculate the term pair index will only consider the minimum distance between term pairs *minDis(t_i,t_j)* as shown by (4). To ensure all the query terms found on the document have an equal opportunity to participate in the proximity scoring query terms will not be limited to a sliding window size.

$$\mathrm{TPi}(t_i,t_j)=\frac{1}{minDis(t_i,t_j)^2} \qquad (4)$$

The *min_span* distance in the span-based proximity approach will be calculated as shown by (5) then transformed to a span distance score using (6).

$$min\_cover=\left(\frac{ut_n{}^{pos}-ut_1{}^{pos}}{nt}\right) \qquad (5)$$

$$Span=\frac{1}{(min\_cover)^2} \qquad (6)$$

Where:

ut_n^{pos} is the last position of a unique query term noted on the text document, ut_1^{pos} is the first position of a unique query term occurrence noted on the document, *nt* is the number of all the unique query terms within the span. Equation (7) shows how the *Span* score is integrated with the aggregated pair-based score to get the overall document term proximity score.

$$\mathrm{TP}(Q,D)=\left(\frac{\left(\sum_{i=1}^{n}\mathrm{TPi}(t_i,t_j)\right)+Span}{n}\right) \qquad (7)$$

Where:

$\mathrm{TPi}(t_i,t_j)$ is the pair-based proximity score between two key terms, *Span* is the span-based proximity score of the document, *n* is the total number of all query terms on document. In previous work by [16] to ensure a proper modeling of a term proximity score one of the conditions is that, the proximity score should decrease as the distance between query terms increases. The proximity score TP(Q,D) arrived at in this paper satisfies that condition. To ensure that the most relevant document are ranked high in users top-*k* a Relevance Score Value denoted as RSV is calculated and used as a ranking feature which is based on cosine score of query vector and document vector integrated with the overall term proximity score TP(Q,D) as shown by (8).

$$\mathrm{RSV}=SIM(\mathrm{Q},\mathrm{D})+TP(\mathrm{Q},\mathrm{D}) \qquad (8)$$

Where *SIM*(Q,D) is the cosine score. The term proximity score will be added to the results of the information retrieval model to improve the relevance of retrieved documents in user top-*k*.

4.0 EXPERIMENT RESULTS AND ANALYSIS

For purpose of comparing performance of the proposed method, TF-IDF was used as the baseline. The experiment needed the use of agricultural dataset and since no standard agricultural dataset was found one was manually built which was composed of 232 RSS feeds collected from different agricultural websites. To determine the performance of the proposed method three groups each composed of two users were randomly selected as test user groups. Queries were created and each group was given the same set of queries. The queries were run on TF-IDF and the proposed method. Documents retrieved from each query run were inspected for relevance and precision for the top 5 documents and top 10 documents retrieved was calculated on results of the two methods obtained by each group.

Table 1: Precision at 5 Documents

	RSS Title and Description	
METHOD	Precision@ 5	Precision@ 10
TF-IDF	0.69	0.54
Proposed	0.78	0.69

According to these results provided precision was calculated for both methods. The proposed approach to term proximity on TF-IDF has a higher precision compared to the baseline TF-IDF at both top 5 and top 10 documents; the difference can bring greater improvement in information retrieval.

5.0 CONCLUSION AND FUTURE WORK

In conclusion the use of TF-IDF and term proximity is an effective way of retrieving relevant information. TF-IDF with the term proximity approach discussed in this paper performs much better than the use of baseline TF-IDF. The results show evidence of the potential power of term proximity and it is quite clear that a proximity scoring function should be included in any information retrieval model to give a significant contribution to improvement of relevance in top-*k* document. In the future studies the framework can be improved to learn the most preferred user information based on user information usage history then factor such information in the ranking feature. It can also be improved to consider query term order of occurrence and synonyms in overall information retrieval and ranking.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my supervisors Professor Waweru Mwangi and Doctor Michael Kimwele for their support and guidance. Also, thanks to all the members of test groups who helped with the experiments.

REFERENCES

[1] Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirement for a Cocitation Similarity Measure, with Special Reference to Pearson’s Correlation Coefficient. Journal of the American Society for Information Science and Technology, 54(6), 550-560.

[2] A. Berger (2000). Bridging the Lexical Chasm: Statistical Approaches to Answer Finding. In Proc. Int. Conf. Research and Development in Information Retrieval, 192-199.

[3] C. Monz. Minimal span weighting retrieval for question answering.In Rob Gaizauskas, Mark Greenwood, and Mark Hepple, editors, Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering, pages 23–30, 2004

[4] C. J. Van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. Journal of Documentation, 33(2):106–119, 1977

[5] D. Hawking and P. Thistlewaite. Proximity operators – so near and yet so far. In Proceedings of the Fourth Text REtrieval Conference (TREC-4), pages 131–143, 1995.

[6] D. Nagao (2008). Web Content Recommender System on RSS using weighted TFIDF University of Aizu, Graduation Thesis. March, 2008.

[7] De Silva, Harsh & Dimuthu Ratnadiwakara, ‘Using ICT to reduce transaction costs in agriculture through better communication: A case-study from Sri Lanka’, mimeo, 2008.

[8] Guo, L. & Peng, Q.K. (2013).A Combinative Similarity Computing Measure for Collaborative Filtering-Applied Mechanics and Materials, Volumes 347-350,pg 2919.

[9] H. Yan, S. Shi, F. Zhang, T. Suel, and J. Wen. E□cient term proximity search with term-pair indexes. In Proceedings of the 19th ACM CIKM, CIKM ’10, pages 1229–1238, 2010.

[10] Jinglei Zhao and Yeogirl Yun. A proximity language model for information retrieval. In SIGIR ’09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in informationretrieval,pages291-298, NewYork, NY, USA, 2009. ACM

[11] Makoto Mukai and Masaki Aono, “A Prototype of Content-based Recommendation System based on RSS,” Tech. Rep. 2005-FI-80, IPSJ SIG, 2005.

[12] R. Schenkel, A. Broschart, S. Hwang, M. Theobald and G. Weikum. Efficient text proximity search. In Proc. of the 14th String Processing and Information Retrieval Symposium, 2007.

[13] R. Song, M. Taylor, J. Wen, H. Hon, Y. Yu. Viewing term proximity from a different perspective. vol 4956, pp. 346357, Springer Berlin /Heidelberg, 2008

[14] R. Cummins and C. O’Riordan. An axiomatic study of learned term-weighting schemes. Learning in a Pairwise Term-Term Proximity Framework for Information Retrieval - SIGIR’09, July 19–23, 2009, Boston, Massachusetts, USA.

[15] S. Buttcher, C. Clarke, and B. Lushman. Term proximity scoring for ad-hoc retrieval on very large text collections. In SIGIR '03: Proceedings of the 26nd annual international ACM SIGIR conference on Research and development in information retrieval, 2006.

[16] T. Tao and C. Zhai. An exploration of proximity measures in information retrieval. In SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 295–302, New York, NY, USA, 2007. ACM.

[17] Y. Lv and C. Zhai. Positional language models for information retrieval. In SIGIR 2009,pages 299–306, Boston, MA, USA, 2009. ACM.

[18] Y. Rasolofo and J. Savoy. Term proximity scoring for keyword-based retrieval systems. In Proceedings of the 25th European Conference on IR Research (ECIR 2003), pages 207–218, 2003.