

# A Novel Context Matching Based Technique for Web Document Retrieval

John Zakos

*School of Information and Comm. Technology  
Griffith University  
Queensland 4215 Australia  
j.zakos@gu.edu.au*

Brijesh Verma

*School of Information Technology  
Central Queensland University  
Queensland 4702 Australia  
b.verma@cqu.edu.au*

## Abstract

*This paper presents a novel context matching technique for the retrieval of web documents. The aim of the technique is to dynamically generate a context-based measure of document term significance during retrieval that can be used as a substitute or co-contributor of the term frequency measure. Unlike term frequency, which relies on a term to occur multiple times within a document to be considered significant, context matching is based on the notion that if a term in a given document occurs in that document in the context of the query, then that term is deemed to be significant. Context matching has the ability to potentially determine a term to be significant even if it occurs only once in a large document. The proposed technique has been implemented and the experiments were conducted using a TREC benchmark database. A comparative analysis shows that context matching significantly improves retrieval effectiveness and outperforms previously published results*

## 1. Introduction

The World Wide Web (WWW) has created many challenges for web document retrieval since it emerged in the early 1990s [1]. Being a repository of billions of web documents, applications of indexing [2], extraction [3], classification [4] and search [5] have all contributed in making the retrieval of web documents a reality. Ultimately though, the aim of a web document retrieval system is, given an inputted query, to accurately and effectively retrieve relevant web documents that contain information to satisfy a user's information need. The calculation of query and document term weights that are used by a retrieval process play a central role in the retrieval effectiveness of a system. Traditionally, term weighting indicators such as term frequency (TF) and inverse document frequency (IDF) [2] have been combined to form the TFIDF measure, which has been popularized through the vector space retrieval model [6]. Recent web document retrieval systems [7-9] employ the use of OKAPI weighting for retrieval. It is essentially a probabilistic term weighting

technique that relies on TF observations for determining the importance of a term occurring in a document.

However, TF is not always the best or most useful indicator of term significance. Quite often, there are relevant documents that contain only a single or a few occurrences of a particular term. Consequently, through TF these terms may never be considered important, even if they are occurring in a relevant document. This is especially the case when infrequently occurring terms appear in large documents containing hundreds or often thousands of terms.

Many researchers have incorporated the use of context in information retrieval. Jing et al [10] used context as a basis of measuring the semantic distances between words. They experimented with their technique on a sub-collection of the TREC-4 corpus and achieved an improvement over the baseline, which was the SMART retrieval system. Billhardt et al [11] proposed a context-based vector space model for information retrieval and observed improvements when utilizing contextual information. The WEBSOM [12] system is an example of another way in which context has been used for web retrieval through the clustering of web documents. IntelliZap [13] and Inquirus [14] are context-based web search engines that require the user to specify some contextual information, pertaining to the inputted query, to be used during retrieval.

Anh et al [15] have recently proposed a non-context term weighting technique based on integral impacts. The idea here is that the mapping of TF weights to an integer scale, based on the TF-based ranking of terms within a document, can result in better term weight representations that provide greater efficiency and effectiveness during retrieval. But because the technique was designed with extreme efficiency for indexing and retrieval as a major goal, the technique has some undesirable characteristics. Firstly, TF floating point weight information is lost during mapping into the integral space. This would result in a number of ranked documents in a retrieved set being assigned the same rank score. Secondly, the technique is based on TF and does not address the problem of being able to identify term significance based on non-TF observations.

To overcome these issues and others, we propose a novel term weighting technique called *context matching* (CM). It does not rely on the number of times a particular term appears in a document to determine whether it is significant or not. The technique exploits query context to generate document term significance in a novel way. With CM, a term occurring infrequently within a large document can potentially be given a high confidence. Vice versa, a term occurring frequently within a small document can potentially be assigned a low confidence. This is a significant characteristic of CM that makes it fundamentally different to TF.

The remainder of paper is organized into 4 further sections. The next section describes the proposed technique. Section 3 presents the experimental setup. Section 4 presents an analysis and discussion of results and in section 5 we conclude and outline the future direction of research.

## 2. Proposed Technique

An overview of the context matching technique as part of the retrieval process is shown below in Figure 1.

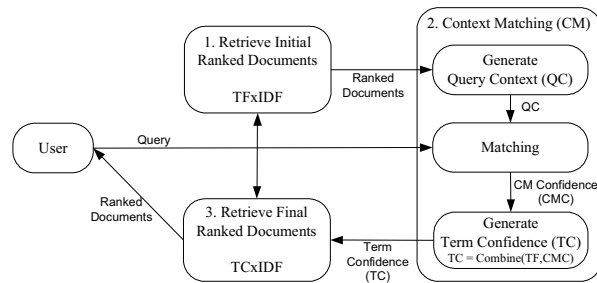


Figure 1. CM as part of the retrieval process

After a query has been inputted, it is used to retrieve an initial set of ranked documents that are passed to the CM technique. The first step of CM is to extract query context from the original query and also from the top ranked documents using a query expansion technique. The query context is then passed to the matching sub-technique that uses it to calculate the term confidence (significance) of the query terms that appear within a document. This term confidence is then combined with TF to give a new term confidence measure that is used to retrieve the final set of documents that are returned to the user. A full description of each main step of CM is given over the following subsections.

### 2.1 Query Context

The context of a query consists of two sub-contexts, each of which are a set of terms: 1. Set of original query terms  $Q$ , and 2. Set of related terms  $QR$ . These

two sets of terms are sub-contexts and together they form the query context  $QC = \{Q, QR\}$ . Each term in each set has a relatedness value in the range  $[0,1]$  that indicates how related that term is to the original query  $Q$ . A value of 1 indicates maximum relatedness where a value of zero indicates that the term not related.

To determine the set of related terms  $QR$  for the query  $Q$ , the technique relies on the use of query expansion using local feedback [16,17]. Typically, an initial run is executed to obtain an initial list of ranked documents and the terms of the top  $n$  documents are assumed to be relevant and then interpreted to select the best  $m$  terms to construct set  $QR$ . We employ the use of

$$TSV_t = w_t r_t \quad (1)$$

to rank candidate terms, where  $w_t$  is a weight (typically IDF) indicating the significance of term  $t$ , and  $r_t$  is the number of relevant documents  $t$  appears in. Terms are ranked using  $TSV_t$  and the top  $m$  are chosen to form  $QR$ . Once  $QR$  has been determined, it is used as part of the query context  $QC$  that can now be used for matching. Each term in  $Q$  and  $QR$  is given a default relatedness value of 1, indicating that it is very related to the original query.  $TSV_t$  could also be used as an interpretation of relatedness.

### 2.2 Matching

The aim of matching is to determine the confidence that a term in a document is relevant to that document in the context of the query. If a query term occurs in a document and if it occurs in the context of the query, then it is considered to be important and given a high confidence. Matching is based on the notion that if two terms co-occur in the same document, then those two terms are related [10] and that the closer together two terms co-occur in a document then the more significant or related they are to each other. Given a term  $q$  and a set of terms that constitute a context  $C$  (i.e.  $Q$  or  $QR$ ), then the *contextual importance* (CI) of the occurrence of  $q$  in document  $D$  can be calculated:

$$CI_{q,C,D} = \frac{\sum_{c \in C} \text{Dist}(CD_{q,c,D}) \times R_c}{\sum_{c \in C} R_c} \quad (2)$$

where  $c$  is a term in the context  $C$ ,  $CD_{q,c,D}$  is the minimum distance between all of the occurrences of  $q$  and  $c$  in  $D$ . Distance is measured by counting the number of words separating  $q$  from  $c$  (this is calculated using term position information stored in the index).  $R_c$  is a value in the range  $[0,1]$  indicating the relatedness of  $c$  to the query (see section 2.1),  $\text{Dist}(CD_{q,c,D})$  is a weighting function of distance importance that returns a value in the range  $[0,1]$ . The smaller  $CD_{q,c,D}$  is the closer

$\text{Dist}(CD_{q,c,D})$  will be to 1. This function can be either of type Gaussian, hard limiter, or linear.

For each query term  $q$ , the technique generates contextual importance using both the original query  $Q$  and related terms  $QR$  as contexts. The final measure is the *context matching confidence* (CMC), which is a combination of the CI from both sub-contexts. Given a query term  $q$  in the query  $Q$ , its CMC is calculated as follows:

$$CMC_{q,D} = (CI_{q,Q,D} \times wI) + (CI_{q,QR,D} \times (1-wI)) \quad (3)$$

where  $wI$  is a weighting factor that is set to 0.5 by default. The resultant CMC is a value in the range [0,1] where a value close to 1 indicates a high confidence that the term  $q$  occurring in document  $D$  is an important indicator of relevance for  $D$  given  $Q$ . A value close to zero indicates low confidence of relevancy. The more related terms that occur at a closer distance to the occurrence of the query term in the document, the higher the resultant confidence. On the other hand, the less related terms that occur at a further distance from the occurrence of the query term, the lower the resultant confidence.

### 2.3 Term Confidence

As mentioned in the introduction, TF-based measures have been used as a useful indicator of term significance by various systems. We chose to incorporate it into CM by combining it with CMC to give a final confidence measure of a term in a document. Given that matching has been performed and we have a CMC value for a term  $q$  in a document  $D$ , the final step of the technique is to combine CMC with TF to give a final term confidence measure. TF is calculated as follows:

$$TF_{q,D} = \frac{\log(count_q + 1)}{\log(numWords_D + 1)} \quad (4)$$

where  $count_q$  is the number of times  $q$  occurred in  $D$ , and  $numWords_D$  is the number of terms in document  $D$ . Having both  $TF_{q,D}$  and  $CMC_{q,D}$ , the term confidence  $TC$  of  $q$  in  $D$  is calculated by the following equation where  $w2$  is a weighting factor that is set to 0.5 by default:

$$TC_{q,D} = (TF_{q,D} \times w2) + (CMC_{q,D} \times (1-w2)) \quad (5)$$

### 3. Experimental Set

The proposed technique has been implemented and experiments were run on the TREC benchmark web document collection WT2g, which consists of 247,491 web documents along with 50 queries with corresponding relevance judgments. A standard inverted index was used to index the collection and each node in the

index stored a document ID, TF and each position of the term in the document. We use up to 2 bytes (16 bits) for term position. Thus, 65535 is the largest term position given to a term in a document. (Any terms in documents exceeding position 65535 are applied a threshold and assigned 65535.)

Given a query  $Q$ , the retrieval function used to calculate the score for document  $D$  is:

$$score_{D,Q} = \sum_{q \in Q} TC_{q,D} \times IDF_q \quad (6)$$

where  $q$  is a term in the query,  $TC_{q,D}$  is the term confidence of term  $q$  in document  $D$  and  $IDF_q$  is the inverse document frequency of  $q$ . This is calculated by:

$$IDF_q = \log \frac{N}{n_q} + 1 \quad (7)$$

where  $N$  is the number of documents in the collection and  $n_q$  is the number of documents in which term  $q$  occurs. To obtain a list of initial documents from which the query expansion technique could extract related terms for  $QR$ , standard TFxIDF was used:

$$score_{D,Q} = \sum_{q \in Q} TF_{q,D} \times IDF_q \quad (8)$$

(This equation is also used to obtain the baseline result for comparison against all other runs.) The top 1000 documents are used for evaluation. We wanted to test the effectiveness of the technique by experimenting with different combinations distance functions  $\text{Dist}(CD)$  and varying the parameters distance  $d$  and number of terms  $m$ . For each experiment,  $d$  was set to 10, 30, 50, 100, 250, 1000 or 65535 and  $m$  was set to 3, 5, 10 or 20.  $\text{Dist}(CD)$  was set to Gaussian, linear or hard limiter. We ran experiments with all 84 combinations of these values for  $d$ ,  $m$  and  $\text{Dist}(CD)$ . For all these experiments,  $n$  was set to 20. This means the top 20 ranked documents of the initial set of retrieved documents were used for query expansion (context generation).

### 4. Results and Analysis

**Table 1. Results of query expansion**

$m$	Avg. Prec.	%	Prec. @ 20	#Rel. Docs
20	0.2918	-2.42%	0.3360	1848
10	0.3066	+2.54%	0.3410	1869
5	0.3364	+12.51%	0.3550	1845
3	0.3303	+10.46%	0.3570	1825

The baseline run, which is standard run that uses Equation 8 for retrieval (without any query expansion or CM), achieved an average precision of 0.2987, a precision at 20 of 0.346 with 1775 relevant retrieved

documents. Table 1 shows the results of runs utilizing traditional query expansion, where expansion terms are added to the original query with down-weighted IDF values and Equation 8 is again used for retrieval. As can be seen, the best traditional query expansion result is when  $m$  is 5 but with only a 12.51% improvement over the baseline. In contrast to traditional query expansion, CM performs remarkably well when using the same expanded terms for  $QR$ . The best 5 results of CM for combinations of parameters selected manually, which uses Equation 6 for retrieval, can be seen in Table 2. The best run which uses  $m = 10$ ,  $d = 250$  and a linear distance function, achieves an average precision of 0.4142, 38.68% better than the baseline run. In fact, all 84 CM runs comfortably outperformed the baseline run and the traditional query expansion runs. The worse performing CM run was when  $m = 3$ ,  $d = 65535$  using a hard limiter distance function. This is not surprising at all as 3 terms do not provide much contextual information and 65535 positions is an extremely large context area to be considering during matching. This run though still achieved an average precision of 0.3428 and outperforms the baseline by 14.75%.

All of the top 5 results shown in Table 2 utilized 5 or more terms for context at a distance of 250 or less. Most of the top results utilized linear or Gaussian functions for distance. This confirms the observation that terms appearing in a close context to each other are a good indicator of significance. The linear and Gaussian functions capture this by rewarding smaller distances with a value closer to 1, whereas the hard limiter distance function ignores this with its constant return value for all values smaller than  $d$ .

To gain an insight into the effect of the different parameters that are combined to calculated term confi-

dence  $TC$ , we ran some further experiments altering  $w1$  and  $w2$ . Table 3 shows the results of experiments for different combinations of  $w1$  and  $w2$ . In the first run,  $w1 = 0$  and  $w2 = 0$ . This in effect is using only the CI calculated using related terms  $QR$  for term confidence (see Equations 3 and 5). No TF information is used here for retrieval. This run yields an impressive average precision of 0.3468. This is 16.10% better than the baseline run which only uses TF. But the number of relevant documents retrieved is only 1525 as compared with the baseline 1775. This is totally understandable though as documents that had no or few occurrences of terms in  $QR$  would not get retrieved or would be ranked outside the top 1000 used for evaluation. Never the less, this run proves the effectiveness of  $CI_{q,D,QR}$  as a reliable indicator of term significance. Run 4 used only the CI calculated from original query terms  $Q$ . This too outperforms the baseline run by 9.28% but like run 1, the number of relevant documents retrieved decreases. Run 3 uses only CMC for retrieval and achieves an average precision of 0.4080. This is extremely encouraging especially since the number of documents retrieved 1756 is close to the number retrieved by the baseline run. Runs 2 and 5, which use TF information as part of the TC calculation, retrieve more documents as expected. CMC indicators seem to be more useful indicators of term significance. Whether used alone or combined with TF, they outperform the baseline run that uses only TF.

The previous best performing system on the same benchmark data was that of Microsoft that achieved an average precision of 0.3829 at TREC 8 [5]. It used OKAPI weighting and traditional query expansion in a content-only retrieval approach. The CM technique outperforms this by 8.1% with its top performing result

**Table 2. Top 5 context matching results**

$m$	$d$	Dist( $CD$ )	Avg. Prec.	%	Prec. @ 20	#Rel. Docs
10	250	Linear	0.4142	+38.68%	0.417	1864
5	100	Linear	0.4137	+38.49%	0.420	1864
5	250	Gaussian	0.4136	+38.47%	0.433	1853
10	100	Linear	0.4135	+38.44%	0.417	1864
10	250	Gaussian	0.4130	+38.27%	0.416	1858

**Table 3. The effect of  $TF_{q,D}$ ,  $CI_{q,Q,D}$  and  $CI_{q,QR,D}$  on retrieval**

#	$w1$	$w2$	Indicators Used	Avg. Prec.	%	Prec. @ 20	# Rel. Docs
1	0	0	$CI_{q,QR,D}$	0.3468	+16.10%	0.3800	1525
2	0	0.5	$CI_{q,QR,D}$ , $TF_{q,D}$	0.3652	+22.27%	0.3870	1870
3	0.5	0	$CI_{q,Q,D}$ , $CI_{q,QR,D}$	0.3980	+33.25%	0.4080	1756
4	1	0	$CI_{q,Q,D}$	0.3264	+9.28%	0.3140	1487
5	1	0.5	$CI_{q,Q,D}$ , $TF_{q,D}$	0.3745	+25.40%	0.3900	1811

of 0.4142, which is a very encouraging result. CM seems to be well suited to the nature of the web information retrieval. This is mainly due to the fact the web queries are typically short (2-4 terms) and most documents that are relevant typically contain at least 1 occurrence of an original query term. This means that context matching can effectively boost these types of documents by matching the context of these original query terms, rather than relying on traditional query expansion.

## 5. Conclusion

We have proposed a novel context matching technique that captures query context and matches this against term contexts in documents to determine term significance and relevancy. The proposed technique has been implemented and tested on the TREC benchmark database. The experimental results showed that the proposed technique has significantly improved the document retrieval performance. It achieved an average precision of 0.4142, which was 8.1% higher than the best results obtained using other existing techniques. In future research, we plan to investigate different combinations of values for  $w1$  and  $w2$  and optimize parameters such as  $\text{Dist}(CD)$ ,  $m$ ,  $n$  and  $d$ .

## Acknowledgement

The research work was partially funded by Central Queensland University's merit research grant scheme.

## References

- [1] World Wide Web Consortium, <http://www.w3c.org>.
- [2] I. Whitten, A. Moffat and T. Bell, "Managing Gigabytes," Morgan Kaufmann Publishers, San Francisco, 1999.
- [3] J. Hu and A. Bagga, "Identifying Story and Preview Images in News Web Pages," in Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, pp. 640-644, 2003.
- [4] A. Schenker, M. Last, H. Bunke, A. Kandel, "Classification of Web Documents Using a Graph Model," in Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, pp. 240-244, 2003.
- [5] S. Robertson and S. Walker, "Okapi/Keenbow at TREC-8," in Proceedings of the 8th Text Retrieval Conference (TREC-8), Gaithersburg, USA, pp. 151-161, 1999.
- [6] G. Salton, C. Yang and A. Wong, "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, no. 11, pp. 613-620, 1975.
- [7] N. Craswell, D. Hawking, T. Upstill, A. McLean, R. Wilkinson and M. Wu, "TREC 12 Web and Interactive Tracks at CSIRO," in Proceedings of the 12th Text Retrieval Conference (TREC-12), Gaithersburg, USA, pp. 193-203, 2003.
- [8] J. Wen, R. Song, D. Cai, K. Zhu, S. Yu, S. Ye and W. Ma, "Microsoft Research Asia of the Web Track of TREC 2003," in Proceedings of the 12th Text Retrieval Conference (TREC-12), Gaithersburg, USA, pp. 408-417, 2003.
- [9] D. Cai, S. Yu, J. Wen and W. Ma, "Block-based Web Search," in Proceedings of the 27th Annual International Conference on Research and development in Information Retrieval, Sheffield, United Kingdom, pp. 456-463, 2004.
- [10] H. Jing and E. Tzoukermann, "Information Retrieval Based on Context Distance and Morphology," in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 90-96, 1999.
- [11] H. Billhardt, D. Borrajo and V. Maojo, "A Context Vector Model for Information Retrieval," Journal of the American Society for Information Science and Technology, vol. 53, no. 3, pp. 236 - 249, 2002.
- [12] T. Honkela, S. Kaski, K. Lagus and T. Kohonen, "WEBSOM - Self-Organizing Maps of Document Collections," in Proceedings of WSOM '97 (Workshop on Self-Organizing Maps), Espoo, Finland, pp. 310-315, 1997.
- [13] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman and E. Ruppim, "Placing Search in Context: The Concept Revisited," in Proceedings of the 10th International World Wide Web Conference, pp. 406-414, 2001.
- [14] E. Glover, S. Lawrence, M. Gordon, W. Birmingham and C. Lee Giles, "Web Search - Your Way," Communications of the ACM, vol. 44, no. 12, pp. 97-102, 2001.
- [15] A. Anh and A. Moffat, "Collection Independent Document Centric Impacts," in Proceedings of the 9th Australasian Document Computing Symposium, Melbourne, Australia, pp. 25-32, 2004.
- [16] J. Xu and B. Croft, "Query Expansion Using Local and Global Document Analysis," in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 4-11, 1996.
- [17] B. Billerbeck and J. Zobel, "Questioning Query Expansion: An Examination of Behaviour and Parameters," in Proceedings of the Australasian Database Conference, Dunedin, New Zealand, vol. 27, pp. 69-76, 2004.