Adaptive Window Size Selection for Proximity Search

Fawaz Alarfaj
School of Computer Science and Electronic Engineering
University of Essex, Colchester, CO4 3SQ, UK
falarf@essex.ac.uk

Term proximity has been successfully used in many entity retrieval searches and enhance the quality of the retrieval systems. In general, the goal of entity searches is to retrieve a ranked list of entities in response to a user's query. The entities could be organisations, products, location, or people. Some of the proximity models that were successful used association discovery in a window of text rather than in a whole document. All current studies have only investigated fixed window sizes; as such, we propose an adaptive window size approach for proximity searches. In this study, we concentrated on a particular type of entity search: expert finding. We used some of the document's attributes such as document length, average sentence length, and number of candidates in the document to adjust the window size of the document. The results of the experiments indicated that considering the document's features when determining the window size did have an effect on the effectiveness of retrieval and provided much better results than a range of baseline approaches using fixed window sizes.

Expert-Finding, Proximity Search, Adaptive Window

1. INTRODUCTION

Traditionally, search engines function by returning a list of documents in response to a user's query; however, the user's information may not be in the form of documents. In fact, users often search for specific things such as people, organisations, or products [Mishne and de Rijke 2006]. These specialised searches lead to the introduction of entity search engines (e.g. product search engines).

In this work, we focused on a special type of entity search: expert-finding. State-of-the-art expert-finding systems typically measure the candidates' knowledge from the text content of top-ranking documents; these are used to derive associations between candidates and search topics based on co-occurrences [Balog et al. 2012]. The co-occurrence of candidate identifiers and query items is considered to provide evidence of expertise. Additionally, the nature and frequency of the co-occurrences is used to estimate the probability of a person being an expert. The first general assumption is that the more often a candidate is found in a document containing many terms describing the topic, the more likely that he/she will be an expert on this topic. The second assumption is that the closer the candidate identifiers are to the query terms, the stronger the association is between them. Under these assumptions, some studies consider the proximity of query terms and candidate identifiers using fixed-size windows. Zhu *et al.* tested 31 window sizes on the W3C collection ranging from 5 to 1100 words. They found the best window size to be around 200 words. They concluded that smaller window sizes could lead to higher precision, but lower recall. On the other hand, large window sizes led to higher recall, but lower precision [Zhu et al. 2009]. In this poster, we introduce the idea of an *adaptive* window size, in which the size of the window is a function of various document features.

2. ADAPTIVE WINDOW SIZE FOR PROXIMITY RANKING

The window size for the proximity function will be determined for each document based on the following features: **Document Length**: According to Miao et al. [2012], it is more likely that more occurrences of a query topic will be found in large documents. It is also more likely to have irrelevant words (noise). Thus, in order to minimise the negative influence of noise, the window size should be relatively smaller as the document gets longer as a relatively small part of a large document could still be bigger than a relatively large part of a small document. **Candidate Frequency**: This

term is used to refer to the number of candidates found in a document. When a document has more occurrences of candidates' evidence, the window size should be relatively larger to accommodate an increased number of occurrences. **Average Sentence Length**: The window size is adjusted in proportion to the average sentence length (in tokens) of the document. We determined window size based on the three features with the following equation:

$$Window\ Size = \frac{\sigma}{3} * (\log(\frac{1}{DocLength}) * \beta_1 + CanFreq * \beta_2 + AvgSentSize * \beta_3) \quad (1)$$

The β weighting factors, which determine each feature's contribution in the equation, have been set empirically, where $\sum_{i=1} \beta_i = 1$. For each dataset ten topics are used for training β variables, thus having a clear distinction between test and training data. After establishing the size of the window, it is applied to every full match for the query found in the document. Then, candidate evidence neighbouring this term is extracted; each term within the window will be given a weight, depending on its distance from the query. The Gaussian kernel function is used to calculate the weight.

3. EXPERIMENTS

To evaluate our approach, we used two collections, the W3C corpus[1] and CSIRO corpus[2]. Both data sets were used to evaluate expert-finding systems at TREC Enterprise tracks. During this investigation, we used the two-stage model for the initial candidate ranking by calculating the probability of the candidate given the query, $P(ca|q)$, as follows:

$$P(ca|q) = \sum_d P(d|q) \cdot P(ca|d, ca) \quad (2)$$

where $P(d|q)$ is the document relevance to the query, which is calculated by the underlying search engine, and $P(ca|d)$ is the candidate's probability given the document. In our baseline, $P(ca|d)$ was calculated using the full document without a proximity function. In all other experiments, we applied Equation (1) to find the optimal window size for the current document. The proximity functions only considered the occurrences within this window of text. Our first baseline was a frequency-based approach. In this baseline, a $TF - IDF$ weighting scheme was used in order to obtain the candidate's importance in a particular document, while at the same time integrating it with the candidate's general importance. To test the effect of each feature of the document separately, we first generated an adaptive window size with only one feature. We have evaluated our method on number of metrics; however, in this poster we report only MAP. In Table 1, we report the best MAP for each feature for both datasets (i.e. W3C and CSIRO). The top MAPs of (0.3454 for W3C and 0.4482 for CSIRO) were achieved using a Gaussian proximity function with an adaptive window size.[3] We found that the difference between our AdaptiveWindow run and the baseline was statistically significant (using paired t-tests on average precision values at $p < 0.05$).

Feature	CanFreq	AvgSentSize	DocLength
W3C	0.2806	0.2798	0.2777
CSIRO	0.3822	0.3427	0.3198

Table 1: *The Mean Average Precision of the Adaptive Window Approach using only a single feature.*

Dataset	Run	MAP
W3C	Baseline	0.1532
	AdaptiveWindow	0.3454
CSIRO	Baseline	0.2067
	AdaptiveWindow	0.4482

Table 2: *The performance of the Adaptive Window Approach.*

4. CONCLUSIONS

In this poster, we introduced the idea of adaptive window size for proximity search. We found that adopting this method results in significant improvements. As for future work, we plan to investigate the effectiveness of using other document features for determining optimal window size. We also plan to test the adaptive window size method on other expert-finding collections and also on other TREC benchmarks.

REFERENCES

Balog, K., Y. Fang, M. de Rijke, P. Serdyukov, and L. Si (2012). Expertise retrieval. *Foundations and Trends in Information Retrieval 6*(2-3), 127–256.

Balog, K., P. Thomas, N. Craswell, I. Soboroff, P. Bailey, and A. De Vries (2008). Overview of the TREC-2008 enterprise track. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, pp. 1425.

Craswell, N., A. de Vries, and I. Soboroff (2005). Overview of the TREC-2005 enterprise track. In *TREC 2005 Conference Notebook*, pp. 199–205.

Miao, J., J. X. Huang, and Z. Ye (2012). Proximity-based Rocchio's model for pseudo relevance. SIGIR '12, Portland, Oregon, pp. 535–544.

Mishne, G. and M. de Rijke (2006). A study of blog search. *Advances in information retrieval 3936*, 289–301.

Zhu, J., D. Song, and S. Rüger (2009, April). Integrating multiple windows and document features for expert finding. *J. Am. Soc. Inf. Sci. Technol. 60*(4), 694–715.

[1] `http://research.microsoft.com/users/nickcr/w3c-summary.html`

[2] `http://es.csiro.au/cerc/`

[3] For comparison, the best run at TREC 2005 reported a MAP value of 0.2749 [Craswell et al. 2005] and for CSIRO at TREC 2008 a MAP value of 0.4490 [Balog et al. 2008]