Time-based Query Performance Predictors

Nattiya Kanhabua
Dept. of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
nattiya@idi.ntnu.no

Kjetil Nørvåg
Dept. of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
noervaag@idi.ntnu.no

ABSTRACT

Query performance prediction is aimed at predicting the retrieval effectiveness that a query will achieve with respect to a particular ranking model. In this paper, we study query performance prediction for a ranking model that explicitly incorporates the time dimension into ranking. Different time-based predictors are proposed as analogous to existing keyword-based predictors. In order to improve predicting performance, we combine different predictors using linear regression and neural networks. Extensive experiments are conducted using queries and relevance judgments obtained by crowdsourcing.

Categories and Subject Descriptors H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval
General Terms Algorithms, Experimentation, Performance
Keywords Query performance prediction, Time-aware ranking

1. INTRODUCTION

In this paper we study the performance prediction of a query explicitly provided with time, for example, the eruptions of volcanoes in Iceland before 2010. The query is explicitly provided with *temporal information needs*, which means that a user wants to know about volcanic events in Iceland during the years before 2010. To determine query performance accurately, both textual and temporal information must be considered. If the query is predicted to perform very poorly, the system can help a user reformulate the query by performing "query suggestion" of *terms* and/or *time* relevant to the query, e.g., a list of all volcanic mountains and time periods of major eruptions in Iceland. On the other hand, if the query is predicted to be sufficiently *good*, it can gain further improvement by document re-ranking with *time-based pseudo-relevance feedback* [6].

2. PREDICTING TEMPORAL QUERY PERFORMANCE

Let q be a temporal query, D be a document collection, T be a set of all temporal expressions in D. N_D is the total number of documents in D and N_T is the number of all distinct temporal expressions in T. Temporal query performance prediction is aimed at predicting the retrieval effectiveness for q. Because q is strongly time-dependent, both the statistics of the document collection D and the set of temporal expressions T must be taken into account. Temporal query performance prediction is defined as $f(q, D, T) \rightarrow [0, 1]$, where f is a prediction function (so-called a predictor) giving a predicted score that can indicate the effectiveness of q. We are only interested in pre-retrieval predictors because they predict query performance independently from a ranking method as opposed to post-retrieval predictors. Temporal expressions and the publication date of a document is represented as a quadruple [1]: (tb_l, tb_u, te_l, te_u) where tb_l and tb_u are the lower bound and upper bound for the begin boundary of a time interval respectively. Similarly, te_l and te_u are the lower bound and upper bound for the end boundary of a time interval. A temporal query q is composed of keywords q_{text} and temporal expressions q_{time}. A document d consists of the textual part d_{text}, i.e., a bag of words, and the temporal part d_{time} composed of the publication date $PubTime(d)$, and temporal expressions mentioned in the document's contents $ContentTime(d)$ or $\{t_1, \ldots t_k\}$.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

3. TIME-BASED PREDICTORS

We propose ten different *time-based* predictors analogous to *keyword-based* predictors, including *T-AvQL*, *T-AvIDF*, *T-MaxIDF*, *T-AvICTF*, *T-SCS*, *T-SumSCQ*, *T-SumVAR*, *T-AvVAR*, *T-AvPMI* and *T-MaxPMI*. The first time-based predictor *T-AvQL* is similar to the average length of a query in [7]. $\textit{T-AvQL} = \frac{1}{|q_{time}|} \sum_{t \in q_{time}} \frac{(tb_l - te_l) + (tb_u - te_u)}{2}$ For example, a query's temporal expression July 2010 is more specific than 21^{st} century, and the first query should perform better than the latter. Hence, the shorter the time span of query, the better it performs. *T-AvIDF* is determines the specificity of q by leveraging document frequencies as done in [4] and can be computed using the INQUERY *idf* formula.

$$\textit{T-AvIDF}(q_{time}) = \frac{1}{|q_{time}|} \sum_{t \in q_{time}} \frac{\log(N_D + 0.5)/df(t)}{\log(N_D + 1)}$$

$df(t)$ is the number of documents containing t. *T-MaxIDF* is the maximum value of *idf* scores. Similar to [4], the averaged inverse collection *time frequency* is measured as $\textit{T-AvICTF} = \frac{1}{|q_{time}|} \sum_{t \in q_{time}} \log \frac{N_T}{tf(t)}$. $tf(t)$ is the total number of occurrences of t in T. The simplified (pre-retrieval) version of Clarity Score [2] is proposed in [4]. We incorporate time into the simplified Clarity Score given as *T-SCS*.

$$\begin{aligned} \textit{T-SCS}(q_{time}) &= \sum_{t \in q_{time}} P(t|q_{time}) \cdot \log \frac{P(t|q_{time})}{P(t)} \\ &\approx \sum_{t \in q_{time}} \frac{1}{|q_{time}|} \cdot \log \frac{1}{|q_{time}|} \cdot \frac{N_T}{tf(t)} \end{aligned}$$

T-SumSCQ is analogous to the summed collection query similarity [8], and it is aimed at capturing the similarity between q_{time} and all temporal expressions in T.

$$T\text{-}SumSCQ(q_{time}) = \sum_{t \in q_{time}} (1 + \ln tf(t)) \cdot \ln(1 + \frac{N_T}{df(t)})$$

The sum of query weight deviation [8] estimates how difficult it is for the retrieval model to rank documents containing query terms by examining *term weights* e.g. TF-IDF. For a temporal query, *temporal weights* will be determined instead of *term weights*. In this paper, we employ the time-aware ranking method TSU [6] to measure *temporal weights*.

$$T\text{-}SumVAR(q_{time}) = \sum_{t \in q_{time}} \sqrt{\frac{1}{|D_t|} \times \sum_{d \in D_t} (TSU(t, PubTime(d)) - \overline{TSU}(t))^2}$$

where D_t are documents containing t and $|D_t|$ is the size of D_t, or $df(t)$. *T-AvVAR* is the averaged value of *T-SumVAR*. Time-based predictors above ignore the relationship between query terms and time. The query tsunami 2004 should perform better than tsunami 2002 because tsunami and 2004 co-occur in a collection more often than by chance, while tsunami and 2002 rarely occur together. PMI is used to determine the relationship between a query term $w \in q_{text}$ and time $t \in q_{time}$ [3]. *T-AvPMI* is the averaged value of all PMI scores. The maximum score *T-MaxPMI* is also considered in a case that the averaged PMI value is low but at least one pair of query term and time has a high PMI.

4. EXPERIMENTS

The New York Times Annotated Corpus is used and 40 queries and judgments from [1]. Queries with *day*, *month* or *year* are grouped into the category "short period" denoted *SP*, and queries with *decade*, *century* as "long period" queries denoted *LP*. There are two retrieval modes: 1) *inclusive* (both query terms and a temporal expression comprise a query q_{text}) and 2) *exclusive* (only query terms constitute q_{text} and a temporal expression is excluded from q_{text}). We use the time-aware ranking method TSU [6] for determining MAP. Parameters of TSU are an exponential decay rate $DecayRate = 0.5$, $\lambda = 0.5$, and $\mu = 6$ months. We use the Weka implementation to model simple linear regression for a single predictor, and linear regression and neural network for combining multiple predictors as done in [5]. The models are trained using cross-validation of 5 folds with 10 repetitions. The averaged values of correlation coefficient and root mean squared error (RMSE) of 5 folds are reported.

Table 1 shows the results of single predictors, where each predictor is statistically tested with the worst performed predictor (as underlined) using paired t-test with $p < 0.05$ (in bold). Because all queries in the dataset associate with one temporal expression, we omit the result of some predictors, e.g., the results of *T-MaxIDF* and *T-AvIDF* are the same, so we only report one of them. *AvQL* and *T-AvICTF* outperform other predictors for "short period", while *MaxIDF*, *SumSCQ* and *T-SumSCQ* perform best for "long period". RMSE shows similar results, that is, *AvQL* and *T-AvICTF* perform best (having the lowest RMSE) for "short period". *T-AvIDF* is the worst predictor for "long period", and its RMSE value unusually too high (=0.65). We found that the predicted scores of *T-AvIDF* for "long period" queries are very small yielding high RMSE values.

Table 2 shows the results of combination methods using linear regression† and neural networks∓. Each combined predictor is statistically tested with that of the best performing single predictors (that is, *AvQL* for "short period" and *T-SumSCQ* for "long period"). Each time-based predictor is combined with its corresponding keyword-based predictor. E.g., *T-AvQL*† denotes the combining of *T-AvQL* and its keyword-based predictor using linear regression. The combination of all predictors is denoted *ALL*. For "short period" and *inclusive*, both *T-AvQL*†, *T-AvQL*∓, *ALL*† and *T-MaxPMI*† outperform the best single predictor significantly. For "long period' the combined methods do not perform well since the correlation coefficient of *T-SumSCQ* is relatively high (though it is negative).

Table 1: Performance of single predictors.

Predictor	Correlation coefficient				RMSE			
	inclusive		*exclusive*		*inclusive*		*exclusive*	
	SP	*LP*	*SP*	*LP*	*SP*	*LP*	*SP*	*LP*
AvQL [7]	**0.36**	0.27	**0.39**	-0.02	**0.28**	**0.23**	**0.29**	**0.25**
AvIDF [2]	**-0.26**	0.04	-0.20	0.12	0.30	**0.24**	**0.29**	**0.24**
MaxIDF [4]	0.04	**-0.27**	-0.16	**-0.27**	0.29	**0.25**	0.30	**0.25**
AvICTF [4]	-0.13	0.19	-0.18	0.24	0.30	**0.22**	**0.29**	**0.23**
SCS [4]	-0.14	0.21	-0.14	0.24	0.30	**0.22**	**0.29**	**0.23**
SumSCQ [8]	-0.09	-0.05	**0.16**	**-0.45**	0.29	**0.24**	**0.29**	**0.24**
SumVAR [8]	-0.20	0.07	**-0.31**	0.19	0.30	**0.22**	0.31	**0.22**
AvVAR [8]	**-0.20**	0.23	**-0.35**	0.00	0.30	**0.23**	**0.30**	**0.23**
AvPMI [3]	**0.29**	-0.05	**0.28**	0.02	0.30	**0.24**	**0.28**	**0.24**
MaxPMI [3]	**0.32**	-0.06	**0.35**	-0.04	0.28	**0.24**	**0.28**	**0.24**
T-AvQL	**0.19**	0.05	**0.19**	0.05	**0.28**	**0.24**	**0.28**	**0.24**
T-AvIDF	**0.27**	-0.05	**0.27**	-0.05	0.29	0.65	**0.29**	0.65
T-AvICTF	**0.35**	0.08	**0.35**	0.08	**0.27**	**0.25**	**0.27**	**0.25**
T-SumSCQ	-0.02	**-0.59**	-0.02	**-0.59**	0.29	**0.32**	**0.29**	**0.32**
T-SumVAR	**0.21**	-0.07	**0.21**	-0.07	0.28	**0.24**	**0.28**	**0.24**
T-AvPMI	**0.15**	0.23	**0.28**	0.20	0.30	**0.22**	**0.27**	**0.23**
T-MaxPMI	0.02	0.08	0.13	0.08	**0.29**	**0.21**	0.27	**0.21**

Table 2: Performance of combined predictors.

Predictor	Correlation coefficient				RMSE			
	inclusive		*exclusive*		*inclusive*		*exclusive*	
	SP	*LP*	*SP*	*LP*	*SP*	*LP*	*SP*	*LP*
T-AvQL†	**0.50**	-0.07	0.33	-0.10	**0.26**	**0.25**	0.28	**0.24**
T-AvIDF†	-0.10	0.01	-0.05	0.01	0.30	**0.23**	0.30	**0.23**
T-AvICTF†	-0.02	-0.22	-0.02	-0.19	0.30	**0.26**	0.29	**0.26**
T-SCS†	-0.02	-0.16	-0.02	-0.19	0.29	**0.25**	0.29	**0.26**
T-SumSCQ†	-0.04	-0.16	-0.04	-0.19	0.29	**0.25**	0.29	0.32
T-SumVAR†	-0.07	-0.08	-0.07	-0.08	0.29	**0.23**	0.29	**0.23**
T-AvVAR†	-0.06	-0.07	-0.04	-0.07	0.30	**0.23**	0.29	**0.23**
T-AvPMI†	-0.10	0.03	0.41	0.03	0.33	**0.24**	**0.27**	**0.23**
T-MaxPMI†	**0.36**	-0.05	0.30	-0.10	**0.26**	**0.23**	0.28	**0.23**
ALL†	**0.43**	-0.04	0.29	-0.11	0.34	0.32	0.33	0.26
T-AvQL∓	**0.47**	0.13	**0.50**	-0.06	0.30	**0.27**	0.30	**0.26**
T-AvIDF∓	-0.02	-0.29	-0.05	-0.29	0.36	**0.27**	0.37	**0.27**
T-AvICTF∓	0.12	-0.17	0.22	0.01	0.33	**0.26**	0.30	**0.29**
T-SCS∓	0.13	-0.09	0.24	-0.07	0.33	**0.26**	0.31	0.30
T-SumSCQ∓	-0.06	-0.09	-0.11	-0.37	0.33	**0.26**	0.34	**0.24**
T-SumVAR∓	-0.09	-0.03	-0.14	0.03	0.35	**0.23**	0.37	**0.24**
T-AvVAR∓	-0.06	-0.05	-0.02	-0.10	0.34	**0.23**	0.35	**0.24**
T-AvPMI∓	0.11	0.16	0.41	0.16	0.36	**0.27**	0.30	**0.25**
T-MaxPMI∓	0.32	0.18	0.50	0.23	0.31	**0.23**	0.29	**0.23**
ALL∓	0.22	-0.09	0.17	0.00	0.38	0.45	0.44	0.42

5. CONCLUSIONS AND FUTURE WORK

To conclude, time-based single predictors outperform the baseline predictors significantly for "short period" queries, and the combined methods outperform single predictors significantly for most cases. Our planned future work are: 1) increase the number of temporal queries used for analysis, 2) consider time uncertainty as an indicator for predicting query performance, and 3) study post-retrieval prediction for temporal search.

6. REFERENCES

[1] K. Berberich, S. Bedathur, O. Alonso, and G. Weikum. A language modeling approach for temporal information needs. In *Proceedings of ECIR'2010*, 2010.

[2] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of SIGIR'2002*, 2002.

[3] C. Hauff, L. Azzopardi, and D. Hiemstra. The combination and evaluation of query performance prediction methods. In *Proceedings of ECIR'2009*, 2009.

[4] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *Proceedings of SPIRE'2004*, 2004.

[5] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25, July 2007.

[6] N. Kanhabua and K. Nørvåg. Determining time of queries for re-ranking search results. In *Proceedings of ECDL'2010*, 2010.

[7] J. Mothe and L. Tanguy. Linguistic features to predict query difficulty. In *Proceedings of SIGIR Workshop*, 2005.

[8] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proceedings of ECIR'2008*, 2008.