Universidad Nacional de Educación a Distancia

Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

Predicción del rendimiento de consultas basado en rankings de documentos y nuevo marco de evaluación.

Tesis Doctoral

Joaquín Pérez Iglesias
Ingeniero Informático

2012

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA
Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

Predicción del rendimiento de consultas basado en rankings de documentos y nuevo marco de evaluación.

Tesis Doctoral

Joaquín Pérez Iglesias

Ingeniero en Informática por la Universidad Rey Juan Carlos de Madrid.

Director:

Lourdes Araujo Serna

Profesor Titular de Universidad del Departamento de Lenguajes y Sistemas
Informáticos
de la Universidad Nacional de Educación a Distancia

Most of the fundamental ideas of science are essentially simple, and may,
as a rule, be expressed in a language comprehensible to everyone.
Albert Einstein.

Capítulo 1

Abstract

This thesis is focused on the field of Query Performance Prediction (QPP). This subject, which is part of Information Retrieval research area, has received increasing attention in the last years. The main purpose of these prediction techniques is to estimate the quality of the document set retrieved when a query is posed to a search system.

Different proposals have been introduced to tackle this problem. They try to exploit specific properties from the query, the document collection or any other information source. This QPP techniques fall within two main approaches: Pre-Retrieval, where the ranking list returned by the search system is ignored and thus this list is not considered, and Post-Retrieval, where the returned documents are analysed to improve the quality of the final estimations.

This thesis introduces a new Post-Retrieval query performance prediction technique. The proposed method is based on the scores assigned by a ranking function to the returned documents. The main hypothesis behind this technique is that a high dispersion along the document scores could imply a high quality of the search system response. This idea raises from some of the ranking functions characteristics, since the computed relevance scores can be observed as quantitative estimations of the documents relevance probability.

The results obtained with this approach are similar to those obtained with other prediction methods considered accurate, such as Clarity Score. In addition, this approach has the advantage of performing the process without using complex techniques, since the predictions are computed by using the standard deviation.

Next, this thesis is focused on the current evaluation framework for prediction methods. Previously, some authors have emphasized some of the main disadvantages that appear with this evaluation framework based on the correlation found between the query predictions and the quality of the queries measured using the relevance judgements.

This evaluation framework shows some of the typical drawbacks due to an evaluation based on correlation. In addition, for the QPP case, measuring the performance of a prediction technique with a correlation coefficient ignores some of its specific properties, since the obtained correlation provides only a measure of the global prediction accuracy. However, for some applications it could be interesting to evaluate the specific predictor performance for high or low quality queries.

In order to avoid the drawbacks mentioned above, a new evaluation framework is introduced in this thesis. This new approach measures the performance of a prediction method classifying different queries by their predicted quality.

Finally, the application of prediction methods to the selective query expansion scenario is also studied. The main purpose of this analysis is to measure the impact of applying these methods within a selective query expansion scenario. That is, the goal is to help a search system to decide in which cases the expansion of a query would improve the quality of the returned documents and thus whether the expansion process should be applied.

This analysis has experimentally proved the inadequacy of the average precision measure as an estimator to decide in which cases a query should be expanded. Besides, it is showed the suitability of predictions based on estimating the P@10 value instead of the average precision obtained by a query in this scenario.

Capítulo 2

Conclusions

This thesis is devoted to the query performance prediction subject. More specifically this work deals with three tasks within the area: prediction methods, prediction evaluation and selective query expansion based on prediction methods estimations.

The first contribution of this work is a new prediction technique proposal. This new approach is focused on computing the topic quality estimations by means of a simple calculation as the standard deviation. This approach provides a similar performance as current methods, while these last employ more complex techniques to compute their estimations.

The second proposal introduced in this thesis is related to the evaluation of query performance prediction methods. In this case, the main purpose is to develop a new evaluation framework able to provide more significant information about the performance of predictions methods, than the supplied by correlation coefficients.

Finally, the quality estimations obtained with prediction methods are applied to the selective query expansion task. The purpose of this study is to measure the potential usefulness of prediction methods within this scenario. Selective query expansion is frequently mentioned as one of the most promising potential applications of query performance prediction methods. However, in the related literature not many examples of a successful integration of these methods have been claimed. Thus, an analysis of the lack of performance in this scenario is carried out in order to achieve a better understanding of this task.

Prediction based on the standard deviation

The main conclusion obtained from the proposed prediction technique described in Chapter 3 is the suitability of standard deviation to predict the performance of a query. These predictions are obtained by simply measuring the standard deviation of the scores assigned by the ranking function to the returned documents, after a query is posed to a search system.

The ability of current ranking functions to distinguish relevant and not relevant documents by means of the assigned scores, supports the use of standard deviation as an estimator of a query performance, since good performing queries will show a higher dispersion among its scores.

Two main aspects can affect the performance of the standard deviation as a prediction method: the number of documents included to measure the standard deviation and the use of a normalization method to standardize the scores assigned by the ranking function. The last is frequently applied in order to facilitate a fair comparison among the scores obtained by different queries.

It can be observed, through the experiments reported in Chapter 3, that the number of documents k included to measure the standard deviation, has a strong effect in relation with the quality of the estimations supplied, consequence of the "long tail" composed of the not relevant documents set. Opposite to this, the application of a normalization process do not cause a significant effect on the computed estimations.

The proposed prediction technique shows a similar performance to other approaches which appear in the related literature such as Clarity Score. The most significant advantage of the proposed method is the capacity of computing estimations without using complex techniques involving a high computational cost. This characteristic of the proposed method makes it suitable for its application to a real search system scenario.

It should be remarked that the query performance prediction technique by means of the standard deviation has caused some impact in the QPP community. Thus, a similar method to the one introduced here was proposed almost simultaneously to our approach (Perez-Iglesias (2009); Pérez-Iglesias and Araujo (2009)) by Shtok et˜al. (2009). Shtok, also proposed the use of the standard deviation as a prediction method, but in this case applying a scores normalization process previously. The normalization factor is based on a statistic computed from the corpus where the prediction method is tested. The results obtained with this approach are similar to the obtained with the proposed method in Chapter **??**.

More recently Ronan Cummins (Cummins et˜al. (2011a)) has proposed an extension to the prediction method introduced in this thesis.

The main difference between both proposals is the method to select the number k of documents included to measure the standard deviation. This k size is computed based on the score received by the first retrieved document. He computes the standard deviation including only those documents with a score higher than half of the score received by the first document in the ranking list. Cummins claims slightly better results than the ones published in this thesis. In a different paper from the same author (Cummins et˜al. (2011b)) and according to their experiments, the authors claim *"The best predictors tend to be the ones based on standard deviations..."*.

Something similar is found by Claudia Hauff, as in her thesis (Hauff (2010), pages 66-67), she remarks the potential of the standard deviation based approach *...retrieval score based methods achieve similar or higher correlations than the evaluated document content based approaches (including Clarity Score).*, with the main advantage of *"the very low complexity, compared to approaches relying on document content, or document and query perturbations."*. As the main drawback of this approach, she remarks *"the reliance on retrieval scores can also be considered a drawback, as such approaches require collaborating search systems that make the retrieval status values available."*.

Proposed evaluation framework

In this thesis a new proposal for the evaluation of query performance prediction methods have been introduced. Although, previous works to this thesis had remarked the necessity of a more suitable approach to evaluate the quality of predictions (Hauff (2010), page 149), no research dealing with this issue has been carried out.

An important part of the work carried out in this thesis is related to the evaluation of query performance prediction methods. Therefore, an analysis dealing with the consequences of an evaluation based uniquely on correlation coefficients, or any other derived measure as Weighted Kendall is done. Based on this study, a new evaluation framework is developed. The main purpose of this new proposal is to avoid some of the drawbacks showed by the current evaluation framework, and thus providing a more informative measure of the QPP methods.

In addition to the well-known issues, which are consequence of the correlation coefficients application as an evaluation measure, described in Chapter 4, it is important to remark that a correlation value only describes the general performance of a prediction method. Thus, it is not possible to measure the accuracy of QPP methods predicting different types of queries according to their performance. In many cases, the possible application of a prediction method to a specific scenario it is not so strongly related to the global performance, but to the performance that a method shows for a specific type of queries.

Based on the drawbacks found, we introduce a new evaluation framework specifically focused on measuring the performance of prediction methods for different type of queries. Thus, this evaluation method is able to show when a prediction method is suitable for the detection of queries with a high performance against predicting queries with a low value of quality. The application of this evaluation framework allows to select the more suitable prediction method for a specific scenario.

The proposed evaluation framework assumes that every query belongs to a class of queries. The class of a query depends on the performance showed by it using measures such as AP or P@10. Since each query is uniquely assigned to a group based on its quality, the evaluation of prediction methods can now be observed as a classification problem and thus applying any measure from this field as accuracy, recall, o F-measure. These measures can be applied to the whole set of queries or to a subset of them, focusing on the performance of the prediction method for a specific type of queries.

As an extension to the classic classification measures, the Distance Based Error Measure (DBEM) measure has been developed. This measure, opposite to others, is focused on the misclassified topics, assigning a different degree of penalty to a wrongly classified topic. The error degree is based on the distance between the real query type and the estimated one. Therefore, with this measure it is possible to observe what type of misclasifications occur more frequently with the evaluated prediction method.

The described evaluation framework is tested with a subset of the current available prediction methods. The performance of these methods is evaluated in a double fashion: by class and for the whole set of topics. Concerning the first evaluation case it is observed experimentally some details about their performance not showed by the correlation coefficients, as the low performance obtained by some of the IDF based prediction methods or the general bias showed by prediction methods to estimate with higher accuracy low quality topics.

In relation to the evaluation for the whole set of topics, a strong correlation ($r = 0{,}7$) between the results obtained with Pearson or Kendall and the global performance obtained with the proposed framework is found. This fact, implies that the proposed framework, when applied to the whole set of topics, shows similar results to the obtained with the classical correlation coefficients.

Selective query expansion

Finally, and based on the results obtained with the previous proposals, the last part of the thesis is devoted to analyze the application of query performance prediction methods to the selective query expansion scenario, when no relevance feedback from the user is available.

This use case is one of the most frequently cited to motivate the potential application of prediction methods to different scenarios. We have concluded experimentally that a low improvement can be expected from the inclusion of a prediction method in a selective query expansion scenario.

This lack of performance is due to the limited relationship between the average precision obtained by a query and the performance of the same query after a query expansion process is applied to it.

This conclusion appears after not observing any improvement when the AP value obtained by a query is applied to decide when it should be expanded to improve its performance. This typical approach of selective query expansion is compared with a random fashion selection of the queries for expansion, obtaining both approaches almost equivalent results.

The same conclusion is observed when a different query expansion method developed specifically for this task is applied. The main characteristic of this query expansion method is the capacity to improve, in all cases, the value obtained after the expansion compared with the average precision obtained originally. This fact suggests that the lack of performance showed in a selective query expansion scenario, using average precision as estimator, is not related to the query expansion method applied.

The conclusions obtained in Chapter 5 contradict the most extended idea about the suitability of the average precision quality measure as an estimator in a selective query expansion scenario. More precisely, the obtained results are opposite to the conclusions obtained by Claudia Hauff (Hauff et˜al. (2010), pages 101-104), related to the correlation threshold necessary to observe an improvement with the application of prediction methods to the selective query expansion scenario.

The contradiction between both results is consequence of two assumptions taken by Hauff. In her experiments Hauff assumes that those topics with an average precision over a fixed threshold will improve their average precision after the expansion process. Simultaneously she assumes that those topics which obtain an average precision lower than another threshold will decrease their average precision when they are expanded.

Finally, the experiments are repeated but using P@10 as estimator for the selection of topics candidates for expansion. In this last case, the obtained results are far from those obtained using average precision as estimator. More important the improvement observed when P@10 is applied as an estimator to select which topics should be expanded it is far from a random selection of topics, contrary to the observed when average precision was used.

Therefore, the obtained results remark the importance of developing prediction methods focused on estimating the suitable evaluation measure, which depends on the scenario where the prediction method is applied.

2.1. Future work

The different prediction methods proposed during the last years have shown a strong capacity to obtain a significant correlation values with the average precision measure. This fact suggests the suitability of these methods to improve the global performance of a search system.

This potential capacity has not been exploited yet to improve user search experience, since a limited improvement has been achieved with the application of query performance prediction methods to real scenarios.

This lack of improvement, despite other causes, could be a consequence of the main objective within this area, which is focused on the estimation of the average precision measure. Although, the average precision is considered as a standard quality measure in information retrieval, this measure is not as relevant as others for specific tasks, as it has been observed in relation to the selective query expansion task.

Future development on new prediction techniques should be guided by the scenario where these methods are applied, and not only by the correlation found between a prediction method and a generic measure of a response quality as average precision.

This change on the performance prediction research methodology will allow the development of new predictors or adaptations of current methods to improve the performance of specific tasks. An example of this new point of view on the development of new prediction techniques appears in the works published by Bellogín (2011) or Bellogín et˜al. (2011). Both works are devoted to the application of prediction techniques to improve the recommendations given to users in a generic recommender system. For this task, the authors adapt the Clarity Score method to this field, and they measure the accuracy of the prediction method in terms of how the recommendations are improved compared to a system where no predictions are applied. The obtained results are very promising and prove the utility of the application of a prediction method to this task.

Based on this new orientation on prediction techniques development, some of the current methods should be adapted in order to predict different quality measures. For instance, it should be tested if the standard deviation can be applied to estimate the quality of a response using a minimum number of documents, as it should be done for P@1 or P@10. In the same way it should be analyzed the performance showed by a method like Clarity Score, when the query language model is built from a small number of documents.

An interesting area of application of prediction methods not studied so far is the automatic generation of relevance judgments. Besides, prediction methods could be combined with different proposal for the automatic generation of relevance judgments, as it was proposed by Soboroff et˜al. (2001). In this work Soboroff, suggests to obtain a set of relevant documents through a random sampling from the set of documents considered relevant by the whole set of systems submitting their results to a TREC task.

Another possible extension focused on improving the accuracy of prediction methods is the combination of different prediction techniques to improve the overall performance. This subject was previously studied by Hauff et˜al. (2009). However, this combination can be carried out more adequately using the information provided by the evaluation framework introduced in Chapter 4. This evaluation makes explicit the weakness and robustness of each prediction technique and thus allowing a more suitable combination.

Predictions methods are, in general, evaluated using standard TREC collections. However, the topics resolved in these collections are mainly informatives. Therefore, an analysis of the performance of predictors based on topic intentionality has not been done. For certain tasks it could be interesting to measure the accuracy of predictions on transactional or navigational topics, opposite to the current framework based on informative topics. Similarly, it could be studied the estimation quality of topics highly ambiguous or topics which are poorly represented along the document collection.

Query performance prediction keeps being an open research question with many fields of application. Some of the current techniques show a significant capacity to provide accurate estimations and these predictions can be employed in many information retrieval related tasks. The right application of these prediction techniques is the main area where the query performance prediction community should be focus on.

Referencias

Alejandro Bellogín, Pablo Castells, and Iván Cantador. Predicting the performance of recommender systems: an information theoretic approach. In *Proceedings of the Third international conference on Advances in information retrieval theory*, ICTIR'11, pages 27–39, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-23317-3.

Alejandro Bellogín. Predicting performance in recommender systems. In Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius, editors, *RecSys*, pages 371–374. ACM, 2011. ISBN 978-1-4503-0683-6.

Ronan Cummins, Joemon Jose, and Colm O'Riordan. Improved query performance prediction using standard deviation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1089–1090, New York, NY, USA, 2011a. ACM. ISBN 978-1-4503-0757-4.

Ronan Cummins, Mounia Lalmas, Colm O'Riordan, and Joemon M. Jose. Navigating the user query space. In *Proceedings of the 18th international conference on String processing and information retrieval*, SPIRE'11, pages 380–385, Berlin, Heidelberg, 2011b. Springer-Verlag. ISBN 978-3-642-24582-4.

C. Hauff. *Predicting the Effectiveness of Queries and Retrieval Systems*. PhD thesis, Univ. of Twente, Enschede, January 2010.

Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. The combination and evaluation of query performance prediction methods. In *ECIR*, pages 301–312, 2009.

Claudia Hauff, Leif Azzopardi, Djoerd Hiemstra, and Franciska de Jong. Query performance prediction: Evaluation contrasted with effectiveness. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan Rüger, and Keith van Rijsbergen, editors, *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*, pages 204–216. Springer Berlin / Heidelberg, 2010. ISBN 978-3-642-12274-3.

Joaquin Perez-Iglesias. Query performance prediction based on ranking list dispersion. In *Third BCS-IRSG Symposium on Future Directions in Information Access (FDIA 2009)*, FDIA 2009. eWIC-BCS, 2009.

Joaquín Pérez-Iglesias and Lourdes Araujo. Ranking list dispersion as a query performance predictor. In *ICTIR '09: Proceedings of the 2nd International Conference on Theory of Information Retrieval*, pages 371–374, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04416-8.

Anna Shtok, Oren Kurland, and David Carmel. Predicting query performance by query-drift estimation. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ICTIR '09, pages 305–312, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04416-8.

Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 66–73, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6.