Active Learning for Learning to Rank

Rodrigo M. Silva[1] ***(Author)***, **Marcos A. Gonçalves**[1] ***(Advisor)***,
Adriano Veloso[1] ***(Co-advisor)***

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
{rmsilva, mgoncalv, adrianov}@dcc.ufmg.br

Abstract. *This paper summarizes our master's dissertation which proposes a novel method for actively sampling query-document instances from a document collection for labeling. Using our technique, it is possible to select and label a small and yet highly effective set that can be used to train Learning to Rank (L2R) algorithms. We conducted extensive experimentation of the method using benchmarking datasets to show that it obtains state-of-the-art results when compared to active and supervised baselines.*

Resumo. *Este artigo resume nossa dissertação de mestrado que propõe um novo método para escolher instâncias de consultas-documentos para serem rotuladas. Usando essa técnica é possível selecionar e rotular um pequeno mas representativo conjunto de exemplos que pode ser usado para produzir modelos de aprendizado de máquina para ordenação de documentos bastante efetivos. Nós conduzimos extensa experimentação do método usando conjuntos de referência para mostrar que ele obtém resultados excelentes quando comparado com outros algoritmos ativos e supervisionados do estado-da-arte.*

1. Context and Motivation

Ranking is an essential feature of many applications. From Web and document search to product recommendation systems and online advertising, results have to be ordered based on their estimated relevance with respect to a query or based on a user profile or preferences. In recent years, there has been an increasing interest in applying Machine Learning techniques to improve ranking performance in a plethora of applications. These Learning to Rank (L2R) algorithms use training sets containing vectors of features that provide information on query-document pairs, plus the assessed relevance of each document to the query, to produce a model or function that relate feature-values to relevance and that can be used to rank the results of new queries. This approach to ranking offers greater flexibility and effectiveness when compared to traditional methods, as it is possible to extend and improve ranking models by adding new features or more instances to the training set.

In order to be able to use a L2R method one usually needs to have large training sets, as the effectiveness of the learned functions may be directly correlated with the amount of supervised training data available. To create these training sets, human annotators must evaluate the documents returned by hundreds or thousands of queries and label them. This process is costly and laborious, as tens or hundreds of thousands of documents have to be inspected. Moreover, human labeling is prone to "noise", especially in repetitive and time-consuming tasks such as subjectively labeling large amounts of data using fine-grained relevance labels (for example, using up to five relevance levels, from "totally irrelevant" to "completely relevant").

Active learning techniques have been proposed to help deal with the labeling problem. The motivation of active learning is that it may be possible to derive highly effective learning functions by carefully selecting and labeling instances that are "informative" to the learning algorithm. Using active learning, we can reduce the cost of producing training sets for L2R algorithms and even improve the effectiveness of the learned functions by avoiding adding "noisy" instances to the training sets. Furthermore, human annotators can spend more time analyzing the relevance of each selected instance, producing better training data. The product of an active learning method is a small and yet highly effective training set that can be used by supervised learning algorithms to rank new user queries.

In the master's dissertation, we propose a novel active learning method for L2R that is both practical and highly effective. Differently from the few active learning algorithms for L2R proposed in the literature, our method empowers the creation of a very small training set from scratch (i.e. directly from an unlabeled set). The method is actually comprised of two distinct but complementary parts: The first, detailed in Chapter 4 of the dissertation, uses association rules to sample an unlabeled set, selecting document instances based on a simple, yet very elegant diversity principle. This technique is highly effective, obtaining very good results and selecting extremely small training sets. Although effective, this method is not easily extended to select more instances if necessary. If, after using it, there's still labeling budget available or for some other reason it is possible to select and label more instances (to improve rank quality, for example), an extension of the method is desirable. In Chapter 5 of the dissertation we propose a round-based second stage Query-By-Committee (QBC) process that allows for the selection and labeling of as many more instances as desired or possible, given the available resources.

We performed extensive experimentation of both stages of the method using the LEarning TO Rank (LETOR) 3.0 benchmarking collection. Using only the first stage, we obtained training sets ranging in size from 1.12% to 2.28% of the unlabeled sets and yet yielding MAP (Mean Average Precision) and NDCG@10 (Normalized Discounted Cumulative Gain @ 10) results that beat a strong active learning baseline on four of the six datasets tested. These initial results also surpassed a well-known supervised method, SVMRank (using the complete training sets), in half of the datasets. Using both stages of the method, with less than 6% of the unlabeled sets selected, yielded results that surpass in most cases (in average, all cases), state-of-the-art supervised algorithms that use the complete training sets, producing some of the best results ever reported for these datasets (e.g., considering the LETOR 3.0 benchmark baselines published by the collection producers). To put these results in perspective, take, for instance, the TD2003 dataset: instead of labeling 30,000 documents, using the first stage of our method it is possible to obtain state-of-the-art results by labeling only 670 documents. Using the second stage and selecting up to 6% of the original set means labeling only 1,800 documents. The results obtained show that it is not only possible to considerably reduce the labeling costs using active learning, but also to sieve out "noise", producing better training sets. In summary, the main contributions of the dissertation are:

- An Active Rule-based Learning to Rank (ARLR) method (stage 1) that can be used to actively select document instances from an unlabeled set without the need of having an initial seed training set. ARLR produces very small and yet effective training sets with the advantage that it naturally stops selecting instances. A paper describing it was published in the *European Conference on Machine Learning*

and Knowledge Discovery in Databases (ECML PKDD), a Qualis A2 conference [Silva et al. 2011].

- A round-based second stage selection method (QBC) that can be used to select more instances for labeling as necessary (stage 2). This part of the dissertation has been accepted for publication as an article in the *Journal of the American Society for Information Science and Technology (JASIST)*, a Qualis A1 journal.

The resulting combined method is very powerful, flexible and applicable to many real-world scenarios. It empowers users to build a L2R training set from scratch and produce a ranking model to effectively rank results from new queries.

During the development of the dissertation, we also used ARLR to enable a parallel implementation of the supervised on-demand association rule method (RLR), with results published in *Web Information Systems Engineering (WISE)* [De Sousa et al. 2012], a Qualis B1 conference. We also successfully adapted the method to perform active sampling in some other pattern recognition tasks, such as author name disambiguation and vandalism detection. These works were published in the *Joint Conference on Digital Libraries (JCDL)*, a Qualis A2 conference [Ferreira et al. 2012] and in *Theory and Practice of Digital Libraries (TPDL)*, a Qualis B1 conference [Sumbana et al. 2012].

2. Stage 1: Active Rule-based Learning to Rank - ARLR

ARLR explores ideas of the supervised association rule L2R algorithm proposed by [Veloso et al. 2008] (which we refer to as RLR, or Rule-based Learning to Rank). RLR works by generating association rules from the training set and using them to infer the relevance level of documents in the test set (see Chapter 2 of the dissertation for details). ARLR, as an active learning method, can be used to produce a small training set from an unlabeled set which can then be used by a supervised learning algorithm such as RLR (or other well known methods such as SVMRank, RankBoost, etc.). The key insight behind ARLR is that the number of association rules generated by the documents in the unlabeled set is an indication of how much information each of these documents share with the current selected (and labeled) training set. Thus, for each unlabeled document in the collection, ARLR generates association rules from the current labeled set and chooses for labeling the document which generates the fewest amount of rules.

More formally, from an unlabeled set $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ we want to select (and label) highly informative documents to compose a new labeled training set $\mathcal{D}$ such that $|\mathcal{D}| \ll |\mathcal{U}|$. Initially, $\mathcal{D}$ is empty and the algorithm cannot extract any rules from it, so it selects from $\mathcal{U}$ the document that shares *the most* feature-values with the other unlabeled documents. This document is labeled and put into $\mathcal{D}$ (but also remains in $\mathcal{U}$). Then, at each round, the algorithm selects the document that demands the fewest rules (i.e., the document in $\mathcal{U}$ for which there are less matching rules), as it shares fewer feature-values with the documents already selected. If only a few rules are extracted for a document u_i, then this is evidence that $\mathcal{D}$ does not contain documents that are similar to u_i, and thus, the information provided by document u_i is not redundant and u_i is a highly informative document given the documents already in $\mathcal{D}$. If $u_i \in \mathcal{U}$ is inserted into $\mathcal{D}$, then the number of rules for documents in $\mathcal{U}$ that share feature-values with u_i will increase. But the number of rules for those documents in $\mathcal{U}$ that do not share any feature-values with u_i will remain unchanged. Therefore, the number of rules extracted for each document in $\mathcal{U}$ can be used

as an approximation of the amount of redundant information between documents already in $\mathcal{D}$ and documents in $\mathcal{U}$. The result is a very small training set based on a *diversity* criterion: the more diverse documents we have in the training set, the more we cover the feature space with the smallest possible amount of documents.

The algorithm stops when all available documents in $\mathcal{U}$ are less informative than any document already inserted into $\mathcal{D}$. This occurs when ARLR selects a document which is already in $\mathcal{D}$. When this happens, ARLR will keep selecting the same document over and over again, and there is no information gain with the inclusion of this document.

3. Stage 2: Expanding the selection using Query-By-Committee - QBC

ARLR selects a very small training set that can be used by a L2R algorithm to build a ranking model and rank new queries. The training set produced is very small but quite effective, as we will see below. In certain situations, it may be desirable to expand the selected training set. Unfortunately, ARLR does not provide for a simple way to keep sampling the unlabeled set. Therefore, we propose a second stage iterative method that allows for the selection of as many more instances for labeling as desired. It uses a query-by-committee approach to select more instances in a round-based fashion. The concept of using a committee of learners to identify "interesting" data instances is well known in classification. The idea is to use an ensemble of models trained using different data to classify an unlabeled set and those instances that the models most disagree about are deemed most informative and selected for labeling.

Our method uses a different approach to QBC, in which separate *algorithms* are used to produce distinct rankings at each round. Thus, we train three algorithms using the same training set (the labeled data available at each round) and rank the remaining of the unlabeled set using these three learners. Then, for *each document* of each query, we calculate a simple metric to determine which documents of that query the learners most disagree in ranking. At each round, we select the first m documents from each query which have the highest value for the disagreement metric described below. To rank the unlabeled sets, we use three algorithms as our committee: SVMRank, RankBoost and Rule-based Learning to Rank (RLR). These algorithms are trained using the labeled set gathered so far and then used to rank the remaining instances in the unlabeled set.

To allow for document-level selection, we propose a simple metric to choose which documents are more diversely ranked by the committee of learners. We use the Coefficient of Variation between the rankings, which is a normalized measure of dispersion defined as σ/μ (standard deviation over the mean). This metric prioritizes documents with smaller ranking variations at the top of the rankings, which is a desired characteristic, since users are usually only interested in the first few results of a ranked list.

4. Experimental Evaluation

We performed extensive experimentation of both ARLR (by itself) and ARLR-QBC (both stages together) using the LETOR 3.0 benchmarking collection. To simulate an active learning scenario, we consider the training sets of the six LETOR web collections as unlabeled sets from which our method selects documents for labeling. Once the instances are selected and labeled (i.e. we have new training sets), the test sets are ranked using a

Table 1. Gains obtained by ARLR-QBC over Donmez and SVMRank (MAP)

	ARLR-QBC vs. Donmez				ARLR-QBC vs. SVMRank			
MAP	AG7%	MG7%	AG%	MG%	AG7%	MG7%	AG%	MG%
TD2003	***10.54***	***26.32*** (0)	3.93	***26.32*** (0)	**5.39**	***10.65*** (4)	2.27	***10.65*** (4)
TD2004	**6.47**	***11.43*** (1)	3.65	***11.43*** (1)	0.44	3.99 (6)	3.38	**6.73** (22)
HP2003	1.69	2.31 (6)	1.55	2.31 (6)	*-0.37*	0.75 (5)	0.44	1.25 (23)
HP2004	0.99	3.71 (0)	2.20	4.30 (19)	4.24	**8.88** (6)	**5.16**	**8.88** (6)
NP2003	2.22	3.64 (6)	1.89	3.64 (6)	*-4.20*	*-2.53* (6)	*-2.38*	*-0.79* (22)
NP2004	*-1.80*	1.71 (3)	*-0.69*	3.22 (8)	1.49	**5.46** (3)	**5.24**	**9.68** (14)

supervised L2R method and ranking metrics calculated. Results for ARLR and ARLR-QBC are presented in Chapters 4 and 5 of the dissertation using RLR and SVMRank as supervised algorithms to test the effectiveness of the actively selected sets. We compare the results obtained with several baselines to show that both ARLR and ARLR-QBC produce training sets that lead to extremely effective ranking models while selecting for labeling very few documents. ARLR selects from 1.12% to 2.28% of the original training sets and yet running RLR using these sets yields results that beat SVMRank using the full training sets on four of the six datasets tested. The extended selected sets obtained by the two-stage method (ARLR-QBC) produces MAP results that beat all twelve supervised baseline algorithms' results published by the LETOR producers in three of the six datasets. Furthermore, we show that ARLR-QBC is significantly better than a strong active learning method for L2R proposed in [Donmez and Carbonell 2008].

Table 1 summarizes the gains obtained by our method compared to Donmez and SVMRank using the full training set. We separate the numbers into two partitions: the average and maximum gains achieved in rounds 0 to 7 (columns AG7% and MG7%, respectively) and the overall (i.e. considering all rounds) average and maximum gains (AG% and MG%). The reason for doing this partitioning is that, although we run ARLR-QBC for 26 rounds, we believe that in most real-world scenarios only a few rounds should be run. By providing the average and max gains obtained at the first 8 rounds (0-7) we want to show that our method converges faster to good results as compared to Donmez and also that it obtains competitive results selecting less than 6% of the original training sets when compared to a strong supervised method using the complete sets (i.e. SVMRank).

Average and Maximum Gains over Donmez: From Table 1 we can see that ARLR-QBC obtains gains over Donmez on all datasets, except NP2004. The improvement is more impressive on the informational datasets (TD2003 and TD2004), where ARLR-QBC has average results on rounds 0-7 that are over 10% better than Donmez on TD2003 and over 6% better on TD2004. The overall average gains are also good, reaching almost 4% on TD2003. The results for the navigational datasets are more modest, but still quit reasonable, with the gain on rounds 0-7 reaching 2.2% on NP2003. Observe from the MG% column that the maximum gain is very often obtained in the initial rounds (the numbers in parentheses indicate at which round the maximum gain was obtained).

Average and Maximum Gains over SVMRank: From the average gain obtained over SVMRank in the first 8 rounds (column AG7% to the right), we can see that ARLR-QBC surpasses SVMRank in four out of six datasets. This means that our method is able to surpass this strong supervised baseline (which uses 100% of the training sets) while selecting and labeling less than 6% of the original training sets. Moreover, the overall average gain (AG% to the right) is positive in five of the six datasets.

Figure 1 shows the comparison of the peak MAP obtained by ARLR-QBC in the first eight rounds (i.e. less than 6% of the unlabeled sets selected), ARLR (i.e. ARLR sets with RLR as ranking method) and three published LETOR baselines: SVMRank, RankBoost and AdaRank-MAP. SVMRank is an obvious choice, since we use it to obtain the results presented for ARLR-QBC. We chose RankBoost and AdaRank-MAP because they are the only two algorithms (out of the twelve baselines published by the LETOR producers) that obtain the highest MAP scores in two datasets each. As we can see, ARLR-QBC obtains better results than ARLR in all datasets, with the exception of NP2003. ARLR-QBC obtains specially good results on the datasets where ARLR did worse: HP2004 and NP2004. These results show that, although ARLR is able to select very small datasets with very good effectiveness, expanding the selection using the QBC second stage is worth the extra labeling cost. We can also see that ARLR-QBC beats the chosen LETOR baselines in TD2003, HP2004 and NP2004. In fact, ARLR-QBC beats all twelve published LETOR baselines on these datasets.

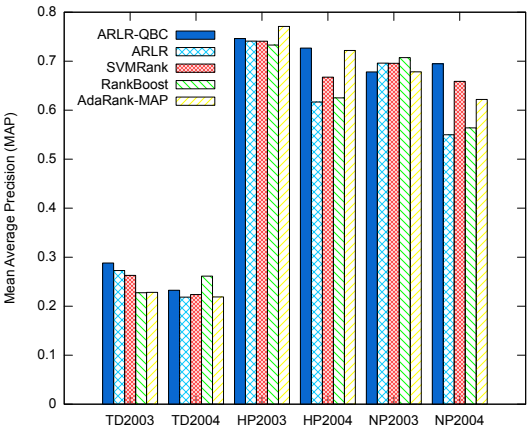

Figure 1. Comparison of ARLR-QBC, ARLR and three LETOR baselines: Peak MAP in rounds 0 to 7

5. Summary

In the master's dissertation we propose a novel two-stage active learning method that is practical, effective and flexible. The method is practical because it facilitates the creation of a small training set for learning to rank, allowing anyone to start using L2R methods on their collections with reduced labeling costs. The resulting training sets are highly effective, providing evidence that carefully selecting the instances to label may reduce "noise" and allow for the creation of high-performing ranking models. Finally, its iterative nature gives the method flexibility, allowing it to be applied in very diverse scenarios and to adapt to different labeling budgets. As an indication of the quality of the dissertation, we published four conference papers (two Qualis A2 and two Qualis B1) and one journal article in the most important journal of the area (Qualis A1, impact factor: 2.081).

References

De Sousa, D. X., Rosa, T. C., Martins, W. S., Silva, R., and Gonçalves, M. A. (2012). Improving on-demand learning to rank through parallelism. In *WISE'12*, pages 526–537.

Donmez, P. and Carbonell, J. G. (2008). Optimizing estimated loss reduction for active sampling in rank learning. In *ICML '08*, pages 248–255.

Ferreira, A. A., Silva, R., Gonçalves, M. A., Veloso, A., and Laender, A. H. (2012). Active associative sampling for author name disambiguation. In *JCDL '12*, pages 175–184.

Silva, R., Gonçalves, M. A., and Veloso, A. (2011). Rule-based active sampling for learning to rank. In *ECML PKDD '11*, pages 240–255.

Sumbana, M., Gonçalves, M. A., Silva, R., Almeida, J., and Veloso, A. (2012). Automatic vandalism detection in wikipedia with active associative classification. In *TPDL'12*, pages 138–143.

Veloso, A. A., Almeida, H. M., Gonçalves, M. A., and Meira, Jr., W. (2008). Learning to rank at query-time using association rules. In *SIGIR '08*, pages 267–274.