Chapter 9

Conclusions and future work

In this thesis we have investigated how to measure and predict the performance of recommender systems. We have analysed and proposed an array of methods based on the adaptation of performance predictors from Information Retrieval – mainly the query clarity predictor, which captures the ambiguity of a query with respect to a given document collection. We have defined several language models according to various probability spaces to capture different aspects of the users and items involved in recommendation tasks. In this context, we have proposed and evaluated novel approaches drawing from Information Theory and Social Graph Theory for different recommender input spaces, using information-theoretic properties of the user's preferences and graph metrics such as PageRank over the user's social network.

Moreover, since we aimed to predict the performance of a particular recommender system, we required a clear recommender evaluation methodology against which performance predictions can be constrasted. Hence, in this thesis we addressed the evaluation methodology as part of the problem, where we have identified statistical biases in the recommendation evaluation – namely the sparsity and popularity biases – which may distort the performance assessments, and therefore may confound the apparent power of performance prediction methods. We have analysed in depth the effect of such biases, and have proposed two experimental designs that are able to neutralise the popularity bias: a percentile-based approach and a uniform-test approach. The systematic analysis of the evaluation methodologies and the new proposed variants have enabled a more complete and precise assessment of the effectiveness of our performance prediction methods.

On the other hand, we have exploited the proposed performance prediction methods in two applications where they are used to dynamically weight different components of a recommender system, namely the dynamic adjustment of weighted hybrid recommendations, and the dynamic weighting of neighbours' preferences in user-based collaborative filtering. Through a series of empirical experiments on several datasets and experimental designs, we have found a correspondence between the predictive power of our performance predictors and performance enhancements in the two tested applications.

In this chapter we present the main conclusions obtained in our research work. In Section 9.1 we provide a summary and a discussion of our contributions, and in Section 9.2 we provide research directions that could be addressed in future work.

9.1 Summary and discussion of contributions

In the next subsections we summarise and discuss the main contributions of this thesis, addressing the research goals stated in Chapter 1. These contributions are organised according to the three main objectives addressed. First, we analysed how to properly evaluate recommender systems in order to obtain unbiased measurements of a recommender system's performance. Second, we proposed performance predictors that aim to estimate the performance of a recommendation method. And third, we used our performance predictors to dynamically combine components of a recommender system.

9.1.1 Analysis of the definition and evaluation of performance in recommender systems

We have analysed different experimental designs existing in the literature about recommender systems, oriented in particular to ranking-based evaluation, and have shown that **assumptions and conditions underlying the Cranfield paradigm are not granted in usual recommendation settings**. Specifically, we have detected statistical confounders (biases) that arise in applying that paradigm to the evaluation of recommender systems. We have shown that the specific value of the evaluation metric has a use for comparative purposes, but has no particular absolute meaning by itself. We have shown that precision decreases linearly with the sparsity of relevant items (**sparsity bias**) in the AR evaluation methodology, whereas it does not suffer from such bias in the 1R approach.

We have also observed that a non-personalised recommender based on item popularity obtains high performance values, and have shown and analysed in detail how this is due to a **popularity bias** in the experimental methodology. To address these issues, we have proposed **novel experimental approaches that effectively neutralise the popularity bias**.

9.1.2 Definitions and adaptations of performance predictors for recommender systems

We have defined and elaborated **performance predictors in the context of recommendation**, usually taking the user as the object of the prediction, but also considering items as an alternative prediction input. Specifically, we have adapted the query performance predictor known as *query clarity* by taking different assumptions and formulations into several variations of **user clarity** predictors. We have also used information theoretical related concepts such as entropy, graph metrics like centrality, PageRank, and HITS, and other domain-specific, heuristic approaches. We have

defined these predictors upon three input spaces of user preferences: **ratings, logs, and social networks**. On ratings and logs we have defined several language models and vocabulary spaces in such a way that our adaptations of clarity would capture different aspects of the user in a unified formulation for both input spaces. Within the same framework, we have introduced the temporal dimension on log-based preference data, drawing and elaborating time-based performance predictors proposed in prior work in the IR field for ad-hoc search.

Additionally, we have defined **item-based predictors** when rating-based preferences are used, which aim to estimate the performance of the items under consideration (to be more precise, the performance of a recommender system in suggesting those items). Here, the main problem is how to define the true performance metric that the predictor is aimed to estimate, since the items are not the main input of the recommendation process. For this reason, we have developed novel methodologies where the performance of an item can be measured, also considering possible biases arising from heavy raters that may distort the results just for statistical reasons.

We have assessed the predictive accuracy of our methods by computing the correlation between estimated and true performance, following standard practice in the IR performance prediction literature. In doing so, we used the unbiased methodologies analysed throughout the thesis to **compare how the predictors behave when the sparsity and popularity biases have been neutralised**. We have found strong correlation values confirming that our approaches result in a **significant predictive power**.

9.1.3 Dynamic weighting in recommender ensembles

Prevalent in the Recommender Systems literature we find combination of recommenders into the so-called recommender ensembles, which are a special type of hybrid recommendation methods where several recommenders are combined, and which are currently very common in the field as represented by current competitions (Bennett and Lanning, 2007; Dror et al., 2012). Collaborative Filtering, one of the major techniques used among the array of available recommendation strategies, can also be seen as a combination of several utility subfunctions, each corresponding to one neighbour (in user-based CF). In the same way performance prediction in Information Retrieval has been used to optimise rank aggregation, we have investigated the use of recommendation performance predictors to dynamically aggregate the output of recommenders and neighbours.

We have defined a **dynamic hybrid framework** where recommender ensembles can benefit from dynamic weights according to performance predictors with which strong correlations have been found. Our results indicate that high correlation with performance tends to correspond with enhancements in dynamic hybrid recommenders. Additionally, dynamic ensembles of recommenders usually outperform

baseline static ensembles for different recommender combinations and the three types of performance predictors investigated.

On the other hand, we have also proposed a **framework for neighbour selection and weighting** in user-based recommender systems. We have defined neighbour performance predictors and metrics by adapting and integrating some of the methods from the trust-aware recommendation literature. Our framework unifies several notions of neighbour performance under the same view, and provides an objective analysis of the predictive power of different neighbour scoring functions. Once the predictive power of these neighbour predictors was confirmed, we used them to weight the information coming from each neighbour in a dynamic fashion, by means of different strategies that combine similarity values and neighbours' weights. Our experiments confirm a correspondence between the correlation analysis and the final performance results, in the sense that the correlation values obtained between neighbour performance predictors and neighbour performance metrics anticipate which predictors will perform better when introduced into the user-based collaborative filtering algorithm.

9.2 Future work

Performance prediction in recommendation is an interesting research topic also from a business perspective, since one could decide when to deliver certain item recommendations to a user, avoiding lowering the user's confidence on the relevance of the recommendations. In this sense, performance predictions of potential recommendations may give control to the service provider; a control that could be used in various ways, such as recommendation combination methods more general than those addressed in this thesis. Regardless of the plausible applications for industry, and beyond the achievements presented throughout the thesis, we envision the following potential future research lines.

The evaluation of recommender systems still is an object of active research in the field, where several questions need more attention, such as the gap between offline and online experiments, and the missing not at random assumption. Nonetheless, in this thesis we have focused our research on aspects related to the prediction of performance, which requires a deeper understanding of the evaluation methodologies used. In this way, we could **extend our analysis of evaluation methodologies to other ranking metrics** such as those based on two rankings (NDPM, and Spearman's and Kendall's correlations) and those adapted from Machine Learning (e.g. AUC). In this way, we may find that one of these metrics is not influenced by any of the confounders described in Chapter 4, or that none of the design alternatives proposed are able to neutralise these effects. As an example of the interest of this topic, recently in (Pradel et al., 2012) the authors analysed the popularity effects over the

AUC metric and found that considering missing data as a form of negative feedback during training may improve performance, although it may also favour popularity-based recommenders over personalised recommendation methods.

Additionally, it would be beneficial for our research to be able to validate the usefulness of the unbiased measurement of performance with **online evaluations**. This would be valuable for a comparative assessment of the offline observations along with a deeper understanding of the extent to which popularity may be or not a noisy signal. Such a user study would help us determine the real benefits (if any) of receiving popular recommendations, since, for instance, by definition these suggestions are not novel, and probably neither serendipitous nor diverse.

In Chapter 6 we have proposed several performance predictors for recommendation based on the same principles as those denoted in IR as pre-retrieval predictors, like the clarity score, where the output of the retrieval engine (or the recommender system in our case) is not used by the predictor. Based on our results, the research possibilities to investigate more performance predictors for recommendation are abundant. In this line, several authors have exploited the **combination of predictors to obtain higher correlation values and stronger predictive power**, such as (Hauff et al., 2009) and (Jones and Diaz, 2007), where penalised regression and linear regression followed by neural network learning were used respectively. In those works the combination of predictors from different nature improved the correlation against the target evaluation metric – i.e., average precision. Thus, we envision the combination of predictors as a worthwhile direction also for recommendation, especially since we have defined predictors based on different inputs that are expected to have low redundancy between them and, when possible, the combination of such predictors may produce higher correlations for different types of inputs. Examples of these combinations may be the mixture of social and temporal dimensions, item-based temporal predictors, and other contextual dimensions not addressed in this thesis.

Moreover, a future investigation could **analyse and adapt to recommender systems post-retrieval performance predictors** defined in the IR literature, such as those based on the analysis of the score distribution from the recommended items to each user. This may provide predictors with stronger correlations and, thus, with more predictive power of the recommenders' performance, as it occurrs in IR where post-retrieval predictors usually obtain higher correlation values than pre-retrieval predictors. The main limitation of this type of predictors is that they cannot be used directly to adapt the output of the recommender, since the complete output – i.e., the ranking – is typically required for the computation of the predictor values. This would require thinking of different applications where this type of predictors could be applied to recommendation

A particular direction worth considering, and also related to Chapter 6, would be the **use of alternative evaluation approaches** beyond correlation metrics, such as those based on clustering the true and estimated performance values (see Section 5.4.2). In our work we have focused on the use of correlation metrics, mainly Pearson's correlation. These metrics have well-known limitations, such as their sensitivity to outliers, and the small (not significative) differences in correlation when a small number of points is used. For this reason, other approaches for assessing the predictive power of the predictors have been proposed. We have to note, however, that the use of a particular evaluation technique should be focused on their application to specific contexts (Pérez-Iglesias and Araujo, 2010); specifically, this requires defining new applications for performance predictors that match the evaluation metric, which we also envision as a potential future work.

Besides, in the same chapter we developed an evaluation methodology to assess the true performance values of the items, in order to evaluate the proposed item predictors. This methodology should be further validated in order to **obtain a fair measure of item performance**, which at this moment is still an open problem. In that way we would be able to define additional item predictors for other input spaces apart from ratings, and improve the predictiveness of the current item performance predictors.

In Chapter 7 we presented experiments regarding the dynamic combination of recommenders in an ensemble. Those experiments were limited to only one performance predictor for a pair of recommenders. We plan to extend these experiments with **ensembles where two predictors are considered** in order to investigate which conditions should be fulfilled by each pair of predictors in order to improve the performance of the ensemble. A related research direction worth of consideration would be the analysis of the sensitivity of the correlation values for which good performance results are obtained in the dynamic hybrid methods. More specifically, we may consider whether it is better to have an overall strong correlation value (in average) or a not very strong average correlation but better estimates for some particular users, where these users would play a significant role in the system such as the power users defined in (Lathia et al., 2008). A study like the one presented in (Hauff et al., 2010) could then be conducted where simulations of predictors with different correlations are evaluated and their effect on the final performance of the ensembles is compared against each other.

Furthermore, another limitation of the experiments presented in Chapter 7 was that the size of ensembles was always two. We aim to consider **ensembles of N recommenders** and, eventually as mentioned above, using one performance predictor for each recommender. This is a natural but non-trivial step towards a generalisation of the proposed framework to larger ensemble recommenders. Alternatively, Machine Learning techniques could be used to learn the best weights to use in the en-

semble in a user and item basis. In this case, a compromise between the computational costs of each technique (machine learning against performance predictors), their predictive power and the tendency to overfitting should be investigated.

Finally, in Chapter 8 we investigated the dynamic neighbour weighting problem using neighbour performance predictors oriented to error-based metrics. The future work related to this chapter could focus on the **adaptation of the neighbour performance metrics used in our approach to ranking-based metrics**, such as precision and recall. As we have already discussed, error metrics are not the best way to measure performance, although they can be considered appropriate in this context since we want to measure the improvement in accuracy of our approaches, along with facilitating comparisons with the state of the art in trust-aware recommendation, where these metrics are prevalent. Therefore, the use of ranking metrics would be a valuable contribution to the field by itself. Furthermore, once a neighbour performance metric based on a ranking metric is provided, we would be able to measure the correlation of the neighbour predictors described in that chapter with such metric, and analyse in detail the predictive power of predictors for ranking metrics. Ideally, we would be able to obtain a predictor with enough predictive power using both types of neighbour performance metrics (based on error and ranking), although this is not easy to grant in general, since each metric is defined to optimise different parameters and concepts.