Economical Evaluation of Recommender Systems: A Proposal

Kevin Roitero, Giuseppe Serra, and Stefano Mizzaro

Dept. of Mathematics, Computer Science, and Physics, University of Udine, Italy
{roitero.kevin}@spes.uniud.it {giuseppe.serra, mizzaro}@uniud.it

Abstract. The evaluation of information retrieval effectiveness by using fewer topics / queries has been studied for some years now: this approach potentially allows to save resources without sacrificing evaluation reliability. We propose to apply it to the evaluation of recommender systems. We describe our proposal and detail what is needed to put it in practice.

1 Introduction

In Information Retrieval (IR), an essential task is the evaluation of the effectiveness of Information Retrieval Systems (IRSs). To support effectiveness evaluation, several initiatives have been established (such as TREC, NTCIR, etc.); they provide the so called test collections, that are composed by: (i) a document collection; (ii) a set of queries (called topics), which are descriptions of information needs; (iii) a set of relevance judgments made by experts for a subset of topic-document pairs, taken as ground-truth. TREC and other initiatives are often competitions: each IRS has to produce a ranked list of documents for each topic; then, the ranked list and the ground truth are compared: the more the system output is similar to the ground truth, the more the system is effective. Effectiveness is measured computing, for each topic, standard evaluation metrics such as Average Precision (AP), Normalized Discounted Cumulative Gain (NDCG), etc. As result of the evaluation process, systems are ranked according to some measure. A common choice is to compute the average value of the metric over the topics. To reduce human effort, and overall evaluation cost, it has been proposed to estimate IRSs effectiveness on the basis of a limited subset of "a few good topics" [2, 4, 6, 8, 10]. We propose to apply that proposal to the domain of Recommender Systems (RecSys). This paper is structured as follows: Section 2 details the state of the art of IR evaluation using few topics; Section 3 describes our proposal; Section 4 details the needed data; and Section 5 concludes.

2 Background: IR Evaluation with Fewer Topics

The test collection based evaluation method can be unfeasible if too high cost and human effort are required. Therefore, reducing the cost of test collections while preserving their reliability is a key challenge for IR. An approach to reduce the human effort is to use fewer topics. Nowadays, it is still not clear how many topics are required. In fact, Sparck Jones and van Rijsbergen [13] conclude that

250 topics are usually acceptable, and 1,000 are usually needed, while Zobel [15] concludes that 25 topics are reasonable good in predicting the effectiveness evaluation of systems on different set of 25 topics. Buckley and Voorhees [3] state that 25 topics are good, but 50 are always better. Webber et al. [14] conclude that 150 topics are needed. Based on statistical analysis, Sakai [11, 12] estimates a minimum number of topics required to preserve the ability to discriminate system effectiveness. To reduce the number of topics, some researchers investigated strategies to identify the best possible choice of an optimal topic subset. A seminal work by Mizzaro and Robertson [6] proposes to exploit Kleinberg's well known HITS algorithm on a matrix, which represents the interactions between systems and topics. Results show that evaluation is affected by topic ease. Roitero et al. [10] extend and generalize the technique and the results.

Guiver et al. [4] and Berto et al. [2] aim at finding the theoretical optimum for the topic selection strategy: they use exhaustive and heuristic search over all possible subsets of topics of a given cardinality, and show that the theoretical optimum subset of "a few good topics" potentially allows a correct evaluation of systems even for rather low cardinalities. More in detail, for each cardinality they find the topic subsets that provide the highest and lowest values of correlation, considering MAP, with respect to the MAP of the full set of topics. Let us consider an example. At cardinality k, they consider all possible 2^k sets of topics (i.e., the columns of the topic-system matrix, see Figure 1(a)); for each subset, they compute the MAP (Mean of APs over systems) value for the matrix with k topics. Finally, they consider the set which maximizes/minimizes the value of correlation between that MAP value and the MAP obtained when considering all topics (i.e., columns). Figure 1(c) shows the correlation between the MAP computed with k topics and the MAP with all the topics using Kendall's τ rank correlation. The three series in the graph represent: the best/worst topic subset (for each cardinality, the topic subset with the highest/lowest value of correlation), and the Average topic subset (the expected correlation that one would get when selecting topics at random). The best series show that it is theoretically possible to evaluate IRSs using fewer topics. This result is extended by Robertson [8] who shows that it is not clear if it is possible to identify subsets of good topics that are general. Roitero and Mizzaro [9] studies if clustering of topics can be exploited to find such subsets of a few good topics.

3 Our Proposal

We propose to apply the above results on finding a few good topics, obtained in IR evaluation, to RecSys. The advantages would be threefold:

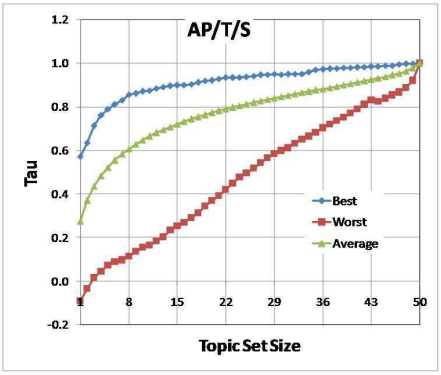
1. We can use our knowledge to evaluate recommender systems saving resources (i.e., using fewer "topics"). Our aim is to obtain a matrix (see Figure 1(b)) having as rows the systems, and as columns the user identifiers (or the recommended items); in the cell we have a "score" assigned by the system to the user (e.g., the satisfaction of the user in being recommended a particular movie, etc.). We can apply the topic reduction approach to reduce the number of users/items required to evaluate the RecSys.

	t_1	$\cdots$	t_n	MAP
s_1	$AP(s_1, t_1)$	$\cdots$	$AP(s_1, t_n)$	$MAP(s_1)$
$\vdots$		$\ddots$		$\vdots$
s_m	$AP(s_m, t_1)$	$\cdots$	$AP(s_m, t_n)$	$MAP(s_m)$

(a)

	u_1	$\cdots$	u_n	AVG
s_1	$E(s_1, u_1)$	$\cdots$	$E(s_1, u_n)$	$AVG(s_1)$
$\vdots$		$\ddots$		$\vdots$
s_m	$E(s_m, u_1)$	$\cdots$	$E(s_m, u_n)$	$AVG(s_m)$

(b)



(c)

Fig. 1: (a) The system–topic matrix: s_i is the i-th system, t_j is the j-th topic, $AP(s_1, t_j)$ is the effectiveness of s_i on t_j; (b) The system–user matrix: s_i is the i-th system, u_j is the j-th user, $E(s_1, u_j)$ is the effectiveness of the system i on the recommendation for the j-th user; (c) Kendall's τ correlation, from [4].

2. We can use our knowledge to build recommender systems while saving resources. Let us consider an example. When implementing a RecSys, a common approach is to obtain a matrix having as rows the users, and as columns the item (e.g., songs, films, etc.). In the matrix cell we have a score assigned by the user to the item (e.g., the times a user listened to a song, an explicit rating, etc.). We can apply the topic reduction approach to reduce the number of items/users required to build the RecSys, preserving its reliability.
3. We can use IR evaluation metrics for the evaluation of recommender systems, and we can transfer the theoretical analysis on such metrics (i.e., robustness, soundness, evaluation effects, etc.) to the metrics currently used in the evaluation of RecSyss. We can adapt IR metrics to RecSys purposes, for example following the categorization of Gunawardana and Shani [5].

4 What We Need

To apply the topic reduction approach to RecSys, we need to obtain a matrix (see Figure 1(b)), which represents the evaluation results and corresponds to the matrix used in IR evaluation (Figure 1(a)). We considered the datasets listed by Özgöbek et al. [7]: for some datasets (The Netflix Prize, MoviePilot Dataset, ACM RecSys Challenge 2016) the data is not public or not available; for some other ones (Movielens Dataset, Million Song Dataset) the dataset contains only the relevance results (i.e., the ground truth). For the RecSyss described by Beel and Dinesh [1], the evaluation data is not publicly available. Finally, for the CLEF NewsReel 2017 dataset,[1] the competition does not look into the individual results of each submission, but instead it gets an aggregated list that

[1] http://www.clef-newsreel.org/

shows the performance of each algorithm on each competition day. Therefore, the datasets we explored so far seem not suitable for our purposes; furthermore, we are not aware of any initiatives for evaluating RecSyss which could be suitable. In absence of suitable data, we might resort to create an artificial recommendation systems collections, starting from "relevance files". We could, for example, sample the relevance files with different relevance probabilities distributions to obtain more/less effective systems; the usage of different relevance distributions with different parameters can provide a realistic population of RecSyss.

5 Conclusions

Summarizing, we propose to: (i) evaluate recommender systems in a more economic way; (ii) reduce the matrix used to build a RecSys; and (iii) use IR evaluation metrics and transfer the metrics properties to the domain of RecSyss. We need some data to experimentally verify the usefulness of our approach.

Bibliography

[1] J. Beel and S. Dinesh. Real-world recommender systems for academia: The pain and gain in building, operating, and researching them [long version]. 2017. arXiv:1704.00156.

[2] A. Berto, S. Mizzaro, and S. Robertson. On using fewer topics in information retrieval evaluations. In *ICTIR*, ICTIR '13, 2013.

[3] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *23rd ACM SIGIR*, pages 33–40. ACM, 2000.

[4] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM TOIS*, 27(4):21, 2009.

[5] A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *J. of Machine Learning Research*, 10(Dec):2935–2962, 2009.

[6] S. Mizzaro and S. Robertson. HITS hits TREC: exploring IR evaluation results with network analysis. In *30th ACM SIGIR*, pages 479–486. ACM, 2007.

[7] Ö. Özgöbek, N. Shabib, and J. A. Gulla. Data sets and news recommendation. In *UMAP Workshops*, 2014.

[8] S. Robertson. On the contributions of topics to system evaluation. In *ECIR*, pages 129–140. Springer, 2011.

[9] K. Roitero and S. Mizzaro. Improving the efficiency of retrieval effectiveness evaluation: Finding a few good topics with clustering? In *Proceedings of the 7th IIR Workshop*, 2016.

[10] K. Roitero, E. Maddalena, and S. Mizzaro. Do easy topics predict effectiveness better than difficult topics? In *ECIR 2017, Aberdeen, Scotland, To be Published*.

[11] T. Sakai. Designing test collections for comparing many systems. In *23rd ACM CIKM*, pages 61–70. ACM, 2014.

[12] T. Sakai. Topic set size design. *Information Retrieval J.*, 19(3):256–283, 2016.

[13] K. Sparck Jones and C. J. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32(1):59–75, 1976.

[14] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *17th ACM CIKM*, pages 571–580. ACM, 2008.

[15] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *21st ACM SIGIR*, pages 307–314. ACM, 1998.