Performance Measures for Multi-Graded Relevance

Christian Scheel, Andreas Lommatzsch, and Sahin Albayrak

Technische Universität Berlin, DAI-Labor, Germany
{christian.scheel,andreas.lommatzsch,sahin.albayrak}@dai-labor.de

Abstract. We extend performance measures commonly used in semantic web applications to be capable of handling multi-graded relevance data. Most of today's recommender social web applications offer the possibility to rate objects with different levels of relevance. Nevertheless most performance measures in Information Retrieval and recommender systems are based on the assumption that retrieved objects (e. g. entities or documents) are either relevant or irrelevant. Hence, thresholds have to be applied to convert multi-graded relevance labels to binary relevance labels. With regard to the necessity of evaluating information retrieval strategies on multi-graded data, we propose an extended version of the performance measure average precision that pays attention to levels of relevance without applying thresholds, but keeping and respecting the detailed relevance information. Furthermore we propose an improvement to the NDCG measure avoiding problems caused by different scales in different datasets.

1 Introduction

Semantic information retrieval systems as well as recommender systems provide documents or entities computed to be relevant according a user profile or an explicit user query. Potentially relevant entities (e. g. users, items, or documents) are generally ranked by the assumed relevance, simplifying user's navigation through presented results. Performance measures evaluate computed rankings based on user-given feedback and thus allow comparing different filtering or recommendation strategies [9].

The most frequently used performance measures in semantic web applications are the *Precision* ($P = \frac{\text{number of relevant items in the result set}}{\text{total number of items in the result set}}$) and the *Mean Average Precision* (MAP) designed to compute the Average Precision over sorted result lists ("rankings"). The main advantage of these measures is that they are simple and very commonly used. The main disadvantage of these measures is, that they only take into account binary relevance ratings and are not able to cope with multi-graded relevance assignments.

One well accepted performance measure designed for handling multi-graded relevance assignments is the *Normalized Discounted Cumulative Gain* (NDCG) [3,8]. From one position in the result list to another the NDCG focuses on the gain of information. Because the information gain of items in the result list

on the same level of relevance is constant, it is possible to swap the positions of items belonging to the same relevance level without changing the performance measure. The advantage of *NDCG* is that it applies an information-theoretic model for considering multiple relevance levels. Unfortunately, the NDCG measure values depend on the number of reference relevance values of the dataset. Thus, NDCG values computed for different datasets cannot be directly be compared with each other.

An alternative point of view to multi-graded relevance was used in the TREC-8 competition [2]. Instead of multiple relevance levels, probabilities for measuring the relevance of entities were used. As performance measure the *Mean Scaled Utility* (SCU) was suggested. Since the SCU is very sensitive to the applied scaling model and the properties of the queries, the SCU measure should not be used for comparing different datasets [2].

Due to the fact, that binary relevance performance measures *precision* and *mean average precision* are commonly used, a promising approach is to extend these binary measures to be capable of handling multi-graded relevance assignments. Kekäläinen et al. [5] discuss the possibility to evaluate retrieval strategies *"on each level of relevance separately"* and then *"find out whether one IR method is better than another at a particular level of relevance"*. Additionally it is proposed to weight different levels of relevance according to their gain of importance. Kekäläinen et al suggest a generalized precision and recall, which contributes to the level of relevance importance, but does not consider the position of an item in the retrieved result list.

In our work we extend the measures *Precision* and *MAP* to be capable of handling multiple relevance levels. The idea of looking at the performance of each level of relevance separately is carried on. An extension of MAP is proposed where strategies can be evaluated with user given feedback independent from the number of used relevance levels. We refer to this extension of MAP as μMAP. Additionally, we introduce an adaptation of the NDCG measure taking into account the number of relevance levels present in the respective reference datasets.

The paper is structured as follows: In the next section we explain the dataset used for benchmarking our work. We explain the performance measure *Average Precision* and show how data has to be transformed in order to compute the *Average Precision*. In Section 3 we propose an extension to *Average Precision* allowing us to handle multi-graded relevance assignment without changing the original ratings. After introducing our approach we evaluate the proposed measures for several Retrieval algorithms and different datasets (Section 4). Finally we discuss the advantages and disadvantages of the new measures and give an outlook to future work.

2 Settings and Methods

For evaluating the performance of a computed item list, a reference ranking is needed (or the items must be rated allowing us to derive a reference ranking). The

reference ranking is expert defined or provided by the user. It can be retrieved explicitly or implicitly [4]. Very popular is the 5-star rating allowing the user to rate entities or items on a scale from 0 to 4, meaning five levels of relevance.

For analyzing and comparing the properties the proposed evaluation measures, we deploy an artificial data set and a real-world dataset, providing three relevance levels. We assume that the reference item ratings stand for ratings coming from human experts and that the test rankings stand for the item list coming from different prediction strategies. We discuss several different types of reference ratings: In each evaluation setting the optimal item list based on the reference ratings should achieve the performance value 1.0.

2.1 Artificial Dataset

We create an artificial dataset and analyze how changes in the dataset influence the measured result quality. For this purpose, we compute the performance of 100 different test item lists for each given reference ranking considering different performance measures.

Test Ratings We create items list ("test rankings") by pair-wise swapping the item of an *optimal* item list ("reference ranking"), see Fig. 1. Swapping means that two rated items in the ranking change their positions. The best test ranking is the one for that no items have been swapped. The performance of the obtained item list decreases with increasing number of swapped item pairs.

The analyzed 100 test rankings differ in the number of the swapped pairs: In the first test ranking (0) we do not swap any item pair, in the last test ranking (99) we randomly swap 99 item pairs. How much the performance decreases per swap depends on the relevance levels' distance of the swapped items. Hence, an evaluation run for each number of switches includes 100 test ranking evaluations to average the results.

Uniformly Distributed Reference Ratings There are four different kinds of reference rankings which differ in the number of relevance levels. Each reference

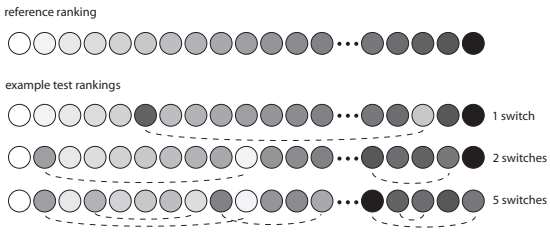

Fig. 1. The figure visualizes the creation of test rankings. Starting with the reference ranking (used for the evaluation) randomly selected item pairs are swapping. The created test rankings differ in the number of swapped pairs.

ranking contains 100 elements which are uniformly distributed among 2, 10, 20, or 50 levels of relevance (see Fig. 2).

Non-Uniformly Distributed Reference Ratings In contrast to the reference rankings used in the previous paragraph, we consider reference rankings consisting of non-uniformly rated items making use of 2, 10, 20, or 50 levels of relevance (see Fig. 3). In other words, the probabilities (that a relevance level is used in the reference ranking) differ randomly between the relevance levels. Moreover, some relevance levels may not be used. Hence, this dataset is more realistic, because users do not assign relevance scores uniformly.

2.2 OHSUMED Dataset

The OHSUMED dataset provided by the Hersh team at Oregon Health Sciences University [1] consists of medical journal articles from the period of 1987–1991 rated by human experts, on a scale of three levels of relevance. Our evaluation is based on the OHSUMED dataset provided in LETOR [6]. The items ("documents") are rated on a scale of 0, 1, and 2, meaning *not relevant*, *possibly relevant* and *definitely relevant*. As in the *Non-Uniformly Distributed Reference Ratings* the given relevance scores are not uniformly distributed.

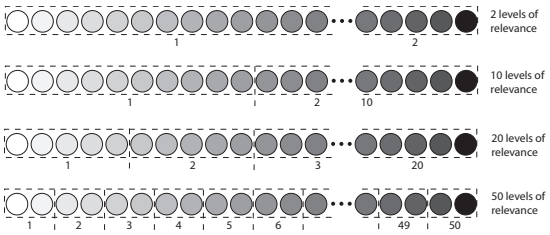

Fig. 2. The Figure visualizes datasets having an almost similar number of items assigned to every relevance level ("uniform distribution of used relevance levels"). The number of relevance levels varies in the shown datasets (2, 10, 20, 50).

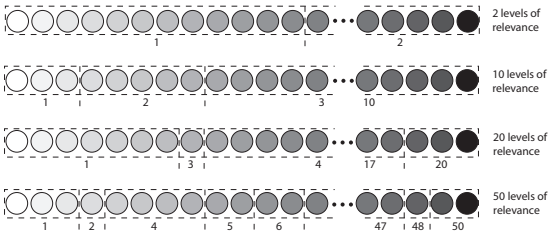

Fig. 3. The Figure shows datasets featuring a high variance in the number of items assigned to a relevance level ("non-uniform distribution of used relevance levels"). There are relevance levels having no items assigned.

Test Ratings The OHSUMED dataset in LETOR provides 106 queries and 25 strategies assigning relevance scores to each item in result set for a respective query. Due to the fact that some the provided strategies show a very similar behavior, we limit the number of evaluated strategies to eight (OHSUMED id 1, 5, 9, 11, 15, 19, 21, 22) enabling a better visualization of the evaluation results.

User-Given Ratings The OHSUMED dataset provides expert-labeled data based on a three level scale. Because there is no real distance between *not relevant*, *possibly relevant* and *definitely relevant*, we assume 1 as distance of successive levels of relevance as the assigned scores 0, 1, and 2 in the dataset imply.

Approximated Virtual-User Ratings The OHSUMED dataset provides three relevance levels. Because fine-grained ratings enable a more precise evaluation, authors believe that soon there will be datasets available with higher number of relevance levels. Until these datasets are available a trick is applied, replacing user's ratings with relevance scores calculated by computer-controlled strategies. The computer calculated relevance scores are treated as "user-given" reference ratings. In our evaluation we selected the OHSUMED strategies *TF of the title* (resulting in 9 different relevance levels) and *TF-IDF of the title* (resulting in 158 different relevance levels) as "virtual" reference users. Both rating strategies show a very strong correlation; the Pearson's correlation coefficient of the relevance assessments is 0.96. The availability of more than three relevance levels in the reference ranking allows us to evaluate ranking strategies with multi-graded relevance assignments. The two strategies treated as reference rating strategies are also considered in the evaluation. Thus, these strategies should reach an evaluation value of 1.

2.3 Performance Measures

There are several performance measures commonly used in information retrieval and recommender systems, such as *precision*, *Area Under an ROC curve* or *rank of the first relevant document (mean reciprocal rank)*. Additionally, the mean of each performance measure over all queries can be computed to overcome the unstable character of some performance measures.

In this section we focus on the popular performance measures *Average Precision* (AP) [10] and *Normalized Discounted Cumulative Gain* (NDCG) [3]). Unlike AP, NDCG can handle different numbers of relevance levels, due to the fact that NDCG defines the information gain based on the relevance score assigned to a document.

Average Precision The average precision of an sorted item ("document") list for a query q is defined as

$$AP_q = \frac{\sum_{p=1}^{N} P@p \cdot rel_q(p)}{R_q} \tag{1}$$

where N denotes the number of the items in the evaluated list, $P@p$ the precision at position p, and R_q the number of relevant items with respect to q. $rel_q(p)$ is a binary function describing if the element at position p is relevant (1) or not (0). A higher AP value means that more relevant items are in the heading of the result list. Given a set of queries, the mean over the AP of all queries is referred to as MAP.

When there are more than two relevance levels, these levels have to be assigned to either 0 or 1. A threshold must be applied, separating the relevant items from the irrelevant items. For later use, we denote AP_q^t as the AP with threshold t applied. AP_q^t is calculated by

$$AP_q^t = \frac{\sum_{p=1}^{N} P@p \cdot rel_q^t(p)}{R_q^t} \quad \text{with} \quad rel_q^t(p) = \begin{cases} 1, & rel_q(p) \geq t \\ 0, & rel_q(p) < t \end{cases} \tag{2}$$

where R_q^t defines the number of results so that $rel_q^t(p)$ is 1.

Normalized Discounted Cumulative Gain For a query q, the normalized discounted cumulative gain at position n is computed

$$\text{NDCG@}n(q) = \text{N}_n^q\text{DCG} = N_n^q \sum_{i=1}^{n} \frac{2^{\text{gain}_q(i)} - 1}{\log(i+1)} \tag{3}$$

where $\text{gain}_q(i)$ denotes the gain of the document at position i of the (sorted) result list. N_n is a normalization constant, scaling the optimal DCG@n to 1. The optimal DCG@n can be retrieved by calculating the DCG@n with the correctly sorted item list.

3 Extending Performance Measures

The need to apply thresholds makes the measures AP and MAP not applicable for multi-graded relevance data. NDCG supports multi-graded relevance data, but the sensitivity to the choice of relevance levels prevents the comparison of NDCG values computed based on different datasets. Hence, for a detailed evaluation based on datasets having multiple relevance levels, both MAP and NDCG have to be adapted.

3.1 Extending Average Precision

In the most commonly used evaluation scenarios, the relevance of items is a binary function (returning "relevant" or "irrelevant"). If the reference dataset provides more than two relevance levels, a threshold is applied which separates the documents into a set of relevant items and a set of irrelevant items. The example in Table 1 illustrates how levels of relevance affect the calculation of the measure AP The example shows a sorted list of items (A ... H). The relevance of

Table 1. The Table shows an example of calculating the average precision for a given item list (each item is rated base on scale of 5 relevance levels). Dependent on the applied threshold t, items are handled as relevant (+) or irrelevant (–). Thus the computed AP depends on the threshold t.

i	A	B	C	D	E	F	G	H	
$rel(i)$	1	0	3	3	2	0	1	4	AP
$t = 5$	–	–	–	–	–	–	–	–	0.000
$t = 4$	–	–	–	–	–	–	–	+	0.125
$t = 3$	–	–	+	+	–	–	–	+	0.403
$t = 2$	–	–	+	+	+	–	–	+	0.483
$t = 1$	+	–	+	+	+	–	+	+	0.780
$t = 0$	+	+	+	+	+	+	+	+	1.000
								mean	0.465
								mean of 5>t>0	0.448

each item is denoted on a scale from 0 to 4 (5 relevance levels). For calculating the precision, a threshold t must be applied, to separate "relevant" from "irrelevant" items. The threshold $t = 0$ implies that all documents are relevant. We refer to this threshold as the irrelevant threshold. In contrast to $t = 0$, applying the threshold $t = 5$ leads to no relevant documents. Table 1 illustrates that the threshold t strongly affects the computed AP. To cope with this problem, we propose calculating the performance on each relevance level, and then computing the mean. This ensures that higher relevance levels are considered more frequently than lower relevance levels. The example visualized in Table 1 shows that item H having a relevance score of 4 is considered relevant more often than all other items.

We refer to this approach as μAP, and μMAP if the mean of μAP for several result lists is calculated. For handling the case that the not all relevance levels are used in every result list and that the "distance" between successive relevance levels is not constant, μAP has to be normalized.

$$\mu AP_q = \frac{1}{\sum_{t \in L} d^t} \sum_{t \in L} \left(AP_q^t \cdot d^t \right) \tag{4}$$

where AP_q^t denotes the average precision using the threshold t, and L a set of all relevance levels (meaning all thresholds) used in the reference ranking. d^t denotes the distance between the relevance level t_i and t_{i-1} if $i > 1$ (and t if $i = 0$). The following example demonstrates the approach: Given a set dataset based on three relevance levels (0.0, 0.3, 1.0), the threshold $t = 0.3$ leads to the $AP^t = 0.3 - 0.0 = 0.3$. The threshold $t = 1.0$ leads to $AP^t = 1.0 - 0.3 = 0.7$.

3.2 The Normalized Discounted Cumulative Normalized Gain

In contrast to MAP, NDCG is designed for handling multiple relevance levels. Unfortunately NDCG does not consider the scale used for the relevance scores.

Table 2. The Table shows how the mapping of relevance scores to relevance levels influences the NDCG measure. In the first example the gain represents an equal match from ratings to relevance levels, in the second example the relevance level is twice the value of the rating, and in the third example the gain of both previous examples is normalized.

	i	A	B	C	D	E	F	G	H	
	rel(i)	1	0	3	3	2	0	1	4	mean
example one:	gain	1	0	3	3	2	0	1	4	
gain equals rel(i)	gain_{opt}	4	3	3	2	1	1	0	0	
	dcg	3.32	3.32	14.95	24.96	28.82	28.82	29.93	45.65	
	dcg_{opt}	49.83	64.50	76.13	80.42	81.70	82.89	82.89	82.89	
	ndcg	0.07	0.05	0.20	0.31	0.35	0.35	0.36	0.55	0.28
example two:	gain	2	0	6	6	4	0	2	8	
gain equals rel(i) $\cdot$ 2	gain_{opt}	8	6	6	4	2	2	0	0	
	dcg	9.97	9.97	114	204	224	224	227	494	
	dcg_{opt}	847	979	1083	1105	1109	1112	1112	1112	
	ndcg	0.01	0.01	0.11	0.19	0.20	0.20	0.20	0.44	0.17
example three:	gain	0.25	0	0.75	0.75	0.5	0	0.25	1	
gain is normalized	gain_{opt}	1	0.75	0.75	0.5	0.25	0.25	0	0	
with ngain (Equ. 5)	dcg	0.63	0.63	1.76	2.74	3.27	3.27	3.48	4.53	
	dcg_{opt}	3.32	4.75	5.88	6.48	6.72	6.94	6.94	6.94	
	ndcg	0.19	0.13	0.30	0.42	0.49	0.47	0.50	0.65	0.39

Thus, computed NDCG values highly depend on the number of relevance levels making it impossible to compare NDCG values between different datasets. Table 2 illustrates this problem.In the first example the NDCG is calculated as usual. In the second example, the number of relevance levels is doubled, but the number of assigned scores as well as the number of used levels of relevance is equal to the first example. This doubling leads to a smaller NDCG compared to the first example, even though no rated element became more or less relevant to another element. In the third example, the gain of example one is normalized and the NDCG is calculated. It can be seen that the normalization solves the inconsistency. A normalization of the gain overcomes the problem of incomparable performance values for data with relevance assignments within a different number of relevance levels. We define the *Normalized Discounted Cumulative Normalized Gain* (NDCNG) at position n as follows:

$$\text{NDCNG@}n(q) = N_n^q \sum_{i=1}^{n} \frac{2^{\text{ngain}_q(i)} - 1}{\log(i+1)}, \ \text{ngain}_q(i) = \begin{cases} \dfrac{\text{gain}_q(i)}{m_q} & , m_q > 0 \\ 0 & , m_q \leq 0 \end{cases} \quad (5)$$

where m_q is the highest reachable gain for the query q ("normalization term"). If there is no relevant item, m_q is set to 0 assuming that irrelevant items are rated with 0. All rating are ≥ 0; relevant items have relevance scores > 0. If these assumptions do not apply, the relevance scores must be shifted so that the irrelevant level is mapped to the relevance score 0.

4 Evaluation

Evaluation based on the Artificial Dataset We evaluated the proposed performance measures on the artificial dataset introduced in Section 2.1. Fig. 4 shows the mean of 100 evaluation runs with uniformly distributed relevance scores. From left to right the number of false pair-wise item preferences increases, and hence the measured performance decreases.

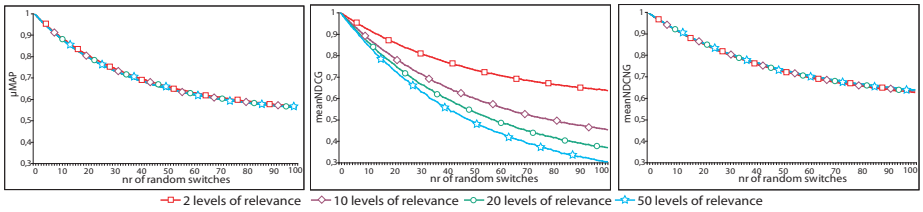

Fig. 4. The evaluation (average over 100 test runs) with the artificial dataset based on uniformly distributed reference ratings shows that in contrast to NDCG, the measures μMAP and NDCNG do not depend on the number of relevance levels.

Fig. 4 shows that in contrast to NDCG, the measures μMAP and NDCNG do not depend on the number of relevance levels. μMAP and the NDCNG calculate the same performance values for similar test rankings. The proposed performance measures explicitly consider the number of relevance levels. This is very important since the common way of applying a threshold to a binary-relevance-based performance measure often leads to a constant performance for item lists differing in the order of items assigned to different relevance levels.

The second evaluation focuses on the analysis how unused relevance levels influence the performance measures. This evaluation is based on the non-uniformly distributed artificial dataset introduced in Section 2.1. Fig. 5 shows that neither μMAP nor NDCNG are affected by the number of items per rank or by the number unused relevance levels.

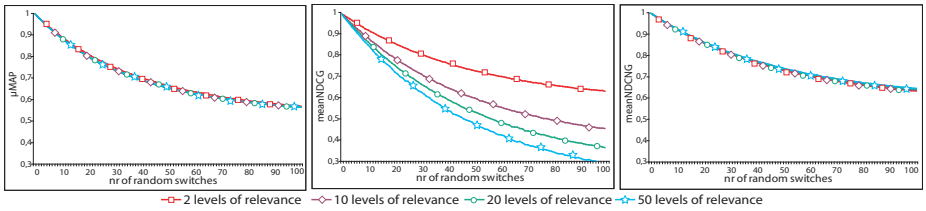

Fig. 5. The evaluation (average over 100 test runs) with the artificial dataset based on a non-uniformly distributed reference ratings shows that NDCG highly depends on the number of relevance levels whereas μMAP and NDCNG do not.

Evaluation based on the OHSUMED Dataset The OHSUMED dataset (introduced in Sec. 2.2) uses three different relevance levels. Fig. 6 visualizes the measured performance of selected retrieval strategies using μMAP, the mean NDCG and the mean NDCNG. Since the OHSUMED dataset uses two relevance levels, the μMAP is the mean of the MAP computed applying the thresholds $t = 1$ and $t = 2$.

A virtual user which is in fact strategy 1 (*TF of title*) provides the relevance assessments for the evaluation presented in Fig. 7. Strategy 1 assigns ordinal scores to 9 different levels of relevance. On this data, μMAP is the mean of 8 different MAP values.

The measured results show, that the measure μMAP evaluates the retrieval strategy 1 (as expected) with a performance value of 1.0, so does the mean NDCG and the mean NDCNG. All the considered evaluation measures agree that the retrieval strategy 9 is most similar to strategy 1, which makes sense, since strategy 9 is computed based on the *TF-IDF of title* and strategy 1 is computed based on *TF of title*. The main difference between both retrieval strategies is the number of used relevance levels: Strategy 1 assigns ordinal relevance scores (using 9 different relevance levels); strategy 9 assigns real values (resulting in 158 relevance levels). The distance between these relevance levels varies a lot.

When applying strategy 9 as reference rating strategy, the need for taking into account the distance between the relevance levels (Equ. 4) can be seen. Several very high relevance scores are used only once; lower relevance scores are used much more frequently. Fig. 8 shows the advantages of the NDCNG compared to "standard" NDCG. The comparison of the mean NDCG in Fig. 7 with the mean NDCG in Fig. 8 reveals that the NDCG is affected by the number of relevance levels. Since the strategies 1 and 9 show a very similar performances in both figures, the other strategies are evaluated with disproportionate lower performance values in Fig. 8 although both reference rating strategies assign similar relevance ratings. The μMAP and the proposed mean NDCNG values do not differ much in both evaluations due to the fact the these measures are almost independent from the number of relevance levels.

5 Conclusion and Discussion

In this paper we introduced the performance measure μAP that is capable to handle more than two levels of relevance. The main advantages of the approach is that it extends the commonly used performance measures *precision* and *Mean Average Precision*. μAP is fully compatible with the "traditional" measures, since it delivers the same performance values if only two reference relevance levels exist in the dataset. The properties of the proposed measures have been analyzed on different datasets. The experimental results show that the proposed measures satisfy the defined requirements and enable the comparison of semantic filtering strategies based on datasets with multi-graded relevance levels. Since μAP is based on well-accepted measures, only a minor adaptation of theses measures is required.

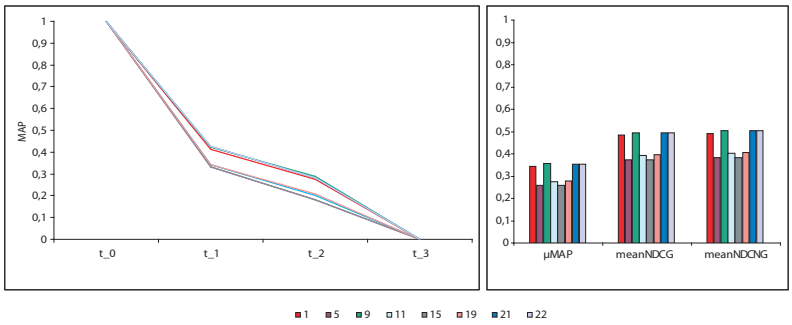

Fig. 6. Performance of selected strategies (OHSUMED id 1, 5, 9, 11, 15, 19, 21, 22). On the left side the mean average precision for each threshold t and on the right side, μMAP, the mean NDCG, and the mean NDCNG value are presented.

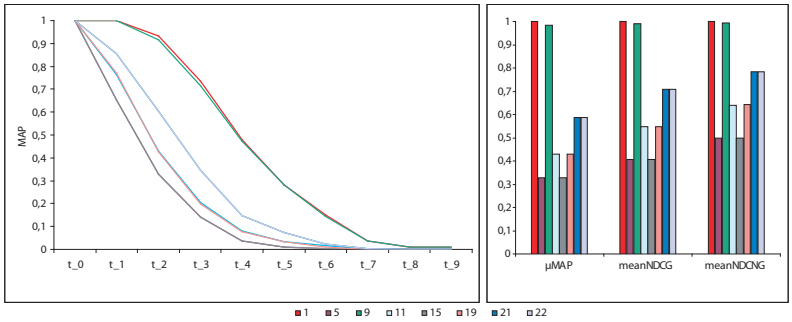

Fig. 7. Performance of selected strategies (OHSUMED id 1, 5, 9, 11, 15, 19, 21, 22), taking strategy TF of title (OHSUMED id 1, 9 levels of relevance) as approximated virtual user.

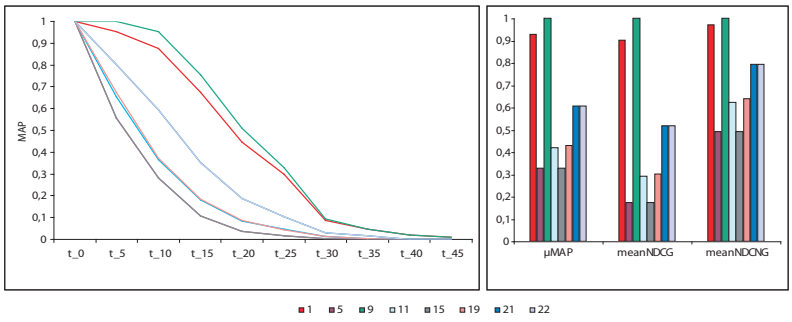

Fig. 8. Performance of selected strategies (OHSUMED id 1,5,9,11,15,19,21,22), taking strategy TF-IDF of title (OHSUMED id 9, 158 levels of relevance) as approximated virtual user.

Additionally, we showed in this paper that the NDCG measure is sensitive to the number of relevance levels in a dataset making it impossible to compare the performance values computed for datasets with a different number of relevance levels. To overcome this problem, we suggest an additional normalization ensuring that the number of relevance levels in the dataset does not influence the computed performance values. Our evaluation shows that NDCNG assigns similar performance values to recommender strategies that are almost similar except that different numbers of relevance levels are used. In the analysis, we demonstrated that high gain values (caused by a high number of relevance levels) lead to incommensurately low NDCG values. Since typically the number of relevance levels differs between the data sets the NDCG values cannot be compared among different data sets. Thus, the gain values per level of relevance must be limited. An additional normalization solves this problem.

Future Work As future work, we plan to use the measures μAP and NDCNG for evaluating recommender algorithms on additional datasets with multi-graded relevance assessments. We will focus on movie datasets such as EACHMOVIE [7] (having user ratings on a discrete scale from zero to five), and movie ratings from the Internet Movie Database (IMDB)[1] (having user ratings on a scale from one to ten).

References

1. W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *SIGIR '94*, pages 192–201, New York, NY, USA, 1994. Springer-Verlag New York, Inc.
2. D. A. Hull and S. Robertson. The trec-8 filtering track final report. In *In The 8th Text Retrieval Conference (TREC-8), NIST SP 500-246*, page 35Ũ56, 2000. http://trec.nist.gov/pubs/trec8/t8_proceedings.html.
3. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.
4. T. Joachims and F. Radlinski. Search engines that learn from implicit feedback. *Computer*, 40(8):34–40, 2007.
5. J. Kekäläinen and K. Järvelin. Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.
6. T.-Y. Liu, T. Qin, J. Xu, W. Xiong, and H. Li. LETOR: Benchmark dataset for research on learning to rank for information retrieval. In *SIGIR Workshop: Learning to Rank for Information Retrieval (LR4IR 2007)*, 2007.
7. P. McJones. Eachmovie collaborative filtering data set. Available from http://research.compaq.com/SRC/eachmovie/, 1997.
8. M. A. Najork, H. Zaragoza, and M. J. Taylor. Hits on the web: how does it compare? In *SIGIR '07*, pages 471–478, New York, NY, USA, 2007. ACM.
9. C. J. Van Rijsbergen. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow, 1979.
10. E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, 2005.

[1] http://www.imdb.com/