**Homework 2: Evaluation Metrics**

Student ID: 18329015 Student Name: 郝裕玮

Date:

Lectured by: Shangsong Liang

Information Retrieval Course

Sun Yat-sen University

**Exercise 1: Rank-based Evaluation Metrics, MAP@K, MRR@K**

**练习 1：基于等级的评估指标，MAP@K，MRR@K**

Assume you have three queries, and the ranking results that a system in response to these three queries are as follows:

Ranking 1 in response to query #1 is: d1, d2, d3, d4, d5, d6, d7, d8, d9, d10. Here only d1, d3, d4, d6, d7, and d10 are relevant (relevance is binary, i.e., either 1 if relevant or 0 if non-relevant) in response to query #1.

Ranking 2 in response to query #2 is: d3, d8, d7, d1, d2, d4, d5, d9, d10, d6. Here only d8 and d9 are relevant in response to query #2.

Ranking 3 in response to query #3 is: d7, d6, d5, d3, d2, d1, d9, d10, d4, d8. Here only d5, d9, and d8 are relevant in response to query #3.

Answer the questions below.

(a) Compute the scores for these metrics: AP@5 (Average Precision @5), AP@10 for each query; RR@5 (Reciprocal Rank score @5), RR@10 for each query.

(b) Compute the scores for these metrics: MAP@5 (Mean Average Precision @5), MAP@10, MRR@5 (Mean Reciprocal Rank score @5), MRR@10 for this system.

假设你有三个查询，系统响应这三个查询的排名结果如下：

响应查询 #1 的排名 1 是：d1、d2、d3、d4、d5、d6、d7、d8、d9、d10。 这里只有 d1、d3、d4、d6、d7 和 d10 是相关的（相关性是二进制的，即，如果相关，则为 1，如果不相关，则为 0）以响应查询 #1。

响应查询 #2 的排名 2 是：d3、d8、d7、d1、d2、d4、d5、d9、d10、d6。 这里只有 d8 和 d9 与查询 #2 的响应相关。

响应查询 #3 的排名 3 是：d7、d6、d5、d3、d2、d1、d9、d10、d4、d8。 这里只有 d5、d9 和 d8 与查询 #3 的响应相关。

回答以下问题。

(a) 计算这些指标的分数：AP@5（平均精度@5），AP@10 用于每个查询； RR@5（Reciprocal Rank score @5），每个查询的 RR@10。

(b) 计算这些指标的分数：MAP@5（平均平均精度@5）、MAP@10、MRR@5（平均倒数排名分数@5）、MRR@10。

**Exercise 2: Rank-based Evaluation Metrics, Precision@K, Recall@K, NDCG@K**

**练习 2：基于等级的评估指标、Precision@K、Recall@K、NDCG@K**

Assume the following ranking for a given query (only results 1-10 are shown); see Table 1. The column ‘rank’ gives the rank of the document. The column ‘docID’ gives the document ID associated with the document at that rank. The column ‘graded relevance’ gives the relevance grade associated with the document (4 = perfect, 3 = excellent, 2 = good, 1 = fair, and 0 = bad). The column ‘binary relevance’ provides two values of relevance (1 = relevant and 0 = non-relevant). The assumption is that anything with a relevance grade of ‘fair’ or better is relevant and that anything with a relevance grade of ‘bad’ is non-relevant.

Also, assume that this query has only 7 documents with a relevance grade of fair or better. All happen to be ranked within the top 10 in this given ranking.

Answer the questions below. P@K (Precision@K), R@K (Recall@K), and average precision (AP) assume binary relevance. For those metrics, use the ‘binary relevance’ column. DCG and NDCG assume graded relevance. For those metrics, use the ‘graded relevance’ column.

假设给定查询的排名如下（仅显示结果 1-10）；见表 1。“rank”列给出了文档的排名。 “docID”列给出了与该级别的文档相关联的文档 ID。 “分级相关性”列给出了与文档关联的相关性等级（4 = 完美，3 = 优秀，2 = 好，1 = 一般，0 = 差）。 “二元相关性”列提供了两个相关性值（1 = 相关，0 = 不相关）。假设是相关等级为“一般”或更好的任何内容都是相关的，而相关等级为“差”的任何内容都是不相关的。

此外，假设此查询只有 7 个相关性等级为公平或更好的文档。在这个给定的排名中，所有这些都恰好排在前 10 名之内。

回答以下问题。 P@K (Precision@K)、R@K (Recall@K) 和平均精度 (AP) 假设二元相关。对于这些指标，请使用“二元相关性”列。 DCG 和 NDCG 假设分级相关。对于这些指标，请使用“分级相关性”列。

Table 1 Top-10 ranking result of a system in response to a query.

表 1 系统响应查询的 Top-10 排名结果。

../../../Writing%20Paper/thinking/fusion/qustion_answering_chaoxin.pdf

(a)  Compute P@5 and P@10.

(a) 计算 P@5 和 P@10。

(b)  Compute R@5 and R@10.

(b) 计算 R@5 和 R@10。

(c)  Provide an example ranking for this query that maximizes P@5.

(c) 为这个查询提供一个最大化 P@5 的示例排名。

(d)  Provide an example ranking for this query that maximizes P@10.

(d) 为这个查询提供一个使 P@10 最大化的示例排名。

(e)  Provide an example ranking for this query that maximizes R@5.

(e) 为这个查询提供一个最大化 R@5 的示例排名。

(f)  Provide an example ranking for this query that maximizes R@10.

(f) 为这个查询提供一个最大化 R@10 的示例排名。

(g)  You have reason to believe that the users of this system will want to examine every relevant document for a given query. In other words, you have recent to believe that users want perfect recall. You want to evaluate based on P@K. Is there a query-specific method for setting the value of K that would be particularly appropriate in this scenario? What is it? (**Hint**: there is an evaluation metric called R-Precision, which we did not talk about in the lectures. Your answer should be related to R-Precision. Wikipedia/Google might help.)

(g) 您有理由相信该系统的用户会希望检查给定查询的每个相关文档。 换句话说，你最近相信用户想要完美的回忆。 您想根据 P@K 进行评估。 是否有一种特定于查询的方法来设置在这种情况下特别合适的 K 值？ 它是什么？ （提示：有一个评估指标叫做 R-Precision，我们在讲座中没有谈到。你的答案应该与 R-Precision 有关。维基百科/谷歌可能会有所帮助。）

(h) Compute average precision (AP). What are the difference between AP and MAP (Mean Average precision)?

(h) 计算平均精度 (AP)。 AP和MAP（平均平均精度）有什么区别？

(i) Provide an example ranking for this query that maximizes average precision (AP).

(i) 为这个查询提供一个最大化平均精度 (AP) 的示例排名。

(j) Compute *DCG*5 (i.e., the discounted cumulative gain at rank 5).

(j) 计算 DCG5（即排名 5 的贴现累积增益）。

(k)  is given by

(k) 由下式给出



where *IDCG*5 is the *DCG*5 associated with the *ideal* top-5 ranking associated with this query. Computing *NDCG*5 requires three steps.

其中 IDCG5 是与此查询相关的理想前 5 排名相关联的 DCG5。 计算 NDCG5 需要三个步骤。

(i) What is the *ideal* top-5 ranking associated with this query (notice that the query has 2 *perfect* documents, 1 *excellent* document, 1 *good* document, 3 *fair* documents, and the rest of the documents are *bad*)?

(i) 与该查询相关的理想 top-5 排名是多少（请注意，该查询有 2 个完美文档、1 个优秀文档、1 个好文档、3 个一般文档，其余文档很差）？

(ii) *IDCG*5 is the *DCG*5 associated with the *ideal* ranking. Compute *IDCG*5. (**Hint:** compute *DCG*5 for your ranking proposed in part (i).)

(ii) IDCG5 是与理想排名相关的 DCG5。 计算 IDCG5。 （提示：为您在第 (i) 部分中提出的排名计算 DCG5。）

(iii) Compute *NDCG*5 using the formula above.

(iii) 使用上述公式计算 NDCG5。

(l) Are there other evaluation metrics to be used to evaluate the performance of the rankings in the table? What are the evaluation scores obtained by these metrics?

(l) 是否有其他评估指标可用于评估表中排名的表现？ 这些指标得到的评价分数是多少？

**Exercise 3: Precision-Recall Curves**

**练习 3：精确召回曲线**

A Precision-Recall (PR) curve expresses precision as a function of recall. Usually, a PR-curve is computed for each query in the evaluation set and then averaged. For simplicity, the goal in this question is to draw a PR-curve for a *single* query. Draw the PR-curve associated with the ranking in Exercise 2 (same query, same results). (**Hint:** Your PR curve should always go down with increasing levels of recall.)

Precision-Recall (PR) 曲线将精度表示为召回率的函数。 通常，为评估集中的每个查询计算 PR 曲线，然后取平均值。 为简单起见，此问题的目标是为单个查询绘制 PR 曲线。 绘制与练习 2 中的排名相关的 PR 曲线（相同的查询，相同的结果）。 （提示：你的 PR 曲线应该随着召回水平的提高而下降。）

**Exercise 4: Other Evaluation Metrics**

**练习 4：其他评估指标**

Except the metrics we have in our lecture slides, are there other evaluation metrics that can be used to evaluate the performance of specific tasks in data mining? What are the tasks and how do to compute such evaluation metrics? (**Hint:** Use the internet to find your answers.)

除了演讲幻灯片中的指标外，还有其他评估指标可用于评估数据挖掘中特定任务的性能吗？ 有哪些任务以及如何计算这些评估指标？ （提示：使用互联网查找您的答案。）