Overall Comparison at the Standard Levels of Recall of Multiple Retrieval Methods with the Friedman Test

José M. Casanova[1], Manuel A. Presedo Quindimil[2], and Álvaro Barreiro[1]

[1] IRLab, Department of Computer Science,
[2] Department of Mathematics,
University of A Coruña,
Campus de Elviña s/n, 15071, A Coruña, Spain
{jcasanova,mpresedo,barreiro}@udc.es,
http://www.dc.fi.udc.es/irlab/

Abstract. We propose a new application of the Friedman statistical test of significance to compare multiple retrieval methods. After measuring the average precision at the eleven standard levels of recall, our application of the Friedman test provides a global comparison of the methods. In some experiments this test provides additional and useful information to decide if methods are different.

1 Introduction

Evaluation is a basic need in Information Retrieval (IR). In order to assess whether a retrieval method performs better than others or not, it is necessary to apply a test of significance, because metrics comparisons are strictly valid for the same collection and queries. The tests of significance determine if the difference is not caused by chance based on statistical evidence. Different applications of these tests have been described [1] and widely used in IR. Evaluation is still an open issue that affects the basis of IR. Recent work assesses that "significance substantially increases the reliability of retrieval effectiveness" [2].

The Friedman test is a non parametric statistical significance test that can be employed with ordinal data [3]. In the ordinary use of this test, it is applied to the Mean Average Precision (MAP) or other metrics using the query as the block variable [4, 5]. We propose to use the Friedman test to compare multiple retrieval methods using the eleven standard levels of recall as the block variable. This new application of the Friedman test provides a global comparison through the levels of recall.

2 The Friedman Test

The Friedman significance test [6] allows the comparison of multiple methods, in situations where the random variable is ordinal (rank-order) and the block variables are mutually independent.

Let b be the number of blocks, k the number of methods to be compared and X the random variable. The function $R(X_{ij})$ returns the rank of method j in the i-th block. In case of tied values, the final rank is the average of the corresponding tied ranking scores. Let $R_j = \sum_{i=1}^{b} R(X_{ij})$ be the sum of ranks for a method. Then the following values A and B are calculated as:

$$A = \sum_{i=1}^{b} \sum_{j=1}^{k} R(X_{ij})^2 \qquad B = \frac{1}{b} \sum_{j=1}^{k} R_j^2 \tag{1}$$

The statistic T is defined as:

$$T = \frac{(b-1)[B - bk(k+1)^2/4]}{A - B} \tag{2}$$

The null hypothesis states that the methods are the same and it is rejected at an α level of significance if the quantile $1-\alpha$ of the F distribution (Fisher-Snedecor distribution) with $(k-1)$ and $(b-1)(k-1)$ degrees of freedom is greater than T.

Paired comparisons among the methods are done when the null hypothesis is rejected. If methods i and j are significantly different the following inequality is satisfied:

$$|R_j - R_i| > t_{1-\alpha/2} \left[\frac{2b(A-B)}{(b-1)(k-1)} \right]^{\frac{1}{2}}, \tag{3}$$

where $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the t-distribution with $(b-1)(k-1)$ degrees of freedom.

3 The Friedman Test with the Standard Levels of Recall as the Block Variable

The Friedman test has been used to determine if differences in MAP and other precision metrics are significant, always using the query as the block variable. We propose a novel application of the Friedman test using as the random variable (X_{ij}) the interpolated average precision at eleven standard levels of recall, i.e. the level of recall acts as the block variable. Intuitively, using the Friedman test as described allows a global comparison of the common precision/recall figures, but with the support of a test of significance. The results obtained with this new test can be different than those obtained when MAP acts as the random and the query as the block variable.

We motivate our evaluation method in the need of an analytical method to compare precision/recall curves. Our approach uses only the information of the ranking position for each method at the 11 standard levels of recall.

The independence assumption regarding the precision at the 11 standard levels of recall may not always hold. If values were dependent, an statistical method for the analysis of repeated measurements [7] could be applied instead of the Friedman test.

4 Experimental Results

To illustrate the use of this test we compared the behaviour of several smoothing methods for language models in the ad-hoc retrieval task. We tried to assess if the following methods described in [8] are different: Jelinek-Mercer (`jm`), Dirichlet (`dir`) and absolute discount (`ad`). The implementation was developed using the retrieval framework Lemur Toolkit[3].

After setting up the best smoothing parameters, we ran the different retrieval methods. Then we computed the MAP and the interpolated average precision at every standard level of recall to evaluate the results. In order to analyse if there was a significant difference among the three smoothing methods we apply the Friedman test to the MAP values and for each of the precisions values at the eleven levels of recall using the query as the block variable. The level of significance of the tests was fixed at $\alpha = 0.05$.

For example, for WEB8 collection (TREC small web) using the titles from topics 401-450 with the smoothing parameters `jm` $\lambda = 0.01$, `ad` $\rho = 0.80$ and `dir` $\mu = 2200$, the Friedman test using the query as the block variable shows that MAP values are significantly different for the three methods, and average interpolated precision values are also significant for the eleven standard levels of recall and the three methods. The Friedman test using the level of recall as the block variable also shows that there are significant differences for the three methods. After doing paired comparisons, the test of MAP values reveals that Dirichlet is better than the other two methods. The tests applied to the precisions at the levels of recall indicate that Dirichlet is significantly better than absolute discounting for all levels of recall and that is better than Jelinek-Mercer for six of the eleven levels. The Friedman test using the recall level as the block variable confirms that Dirichlet is significantly better than the other methods.

Now we describe another scenario where the proposal presented in this paper complements the information provided by previous tests. For the FBIS collection (TREC disk 5) using only the titles from topics 351-400 with the smoothing parameters `jm` $\lambda = 0.05$, `ad` $\rho = 0.75$ and `dir` $\mu = 4000$, the Friedman test using the query as the block variable only finds significant differences among the three methods for precision values at levels of recall from 0.40 to 0.80 and at level 1.00, but it does not find significantly different MAP values. Paired comparisons support that Dirichlet performs better than the other two methods. The Friedman test using the level of recall as the block variable indicates that the three methods are significantly different, and paired comparisons also show that Dirichlet performs better than the others. Therefore, in this experiment the use of the recall level as the block variable for the Friedman test reinforces that Dirichlet is the smoothing algorithm that outperforms all others.

Another interesting experiment is the following. For the LATIMES collection (TREC disk 5) using the title, description and narrative from topics 351-400 with the smoothing parameters `jm` $\lambda = 0.80$, `ad` $\rho = 0.75$ and `dir` $\mu = 3000$, the Friedman test using the query as the block variable shows that MAP values

[3] http://www.lemurproject.org

are no significantly different for the three methods, and average interpolated precision values are only significant at the levels of recall of 0.30, 0.70 and 0.80. The Friedman test using the level of recall as the block variable also shows that there are not statistically significant differences among the three methods.

5 Conclusions and Further Work

We proposed a new application of the Friedman significance test using the levels of recall as the block variable. The use of this variant gives additional information to previous applications of the Friedman test that used the query as the block variable. This way the test provides an approximation to the comparison of the precision/recall curves. We illustrated the method, giving evidence that in fact in some experiments the proposed test helps to decide whether the methods are different or not. We plan to apply our method using the collection as the block variable and the MAP or other single measure as the random variable because this ensures the independence assumption. An alternative to our method for the cases where independence may not hold could be the application of statistical methods for the analysis of repeated measurements. Future work will also address the problem of determining the utility of the test of significance following the methodology introduced by Zobel in [9].

Acknowledgements. The work reported here was cofunded by the SEUI and FEDER funds under the project MEC TIN2005-08521-C02-02 and "Xunta de Galicia" under project PGIDIT06PXIC10501PN. José M. Casanova is supported by a grant of DXID of the "Xunta de Galicia". We also thank the support of the "Galician Network of NLP & IR" Ref. 2006/03.

References

1. Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: Proc. of ACM SIGIR '93. (1993) 329–338
2. Sanderson, M., Zobel, J.: Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proc. of ACM SIGIR '05. (2005) 162–169
3. Sheskin, D.: Handbook of parametric and nonparametric statistical procedures. Chapman & Hall/CRC (2000)
4. Hull, D.: Stemming algorithms: A case study for detailed evaluation. JASIS **47**(1) (1996) 70–84
5. Kekäläinen, J., Järvelin, K.: The impact of query structure and query expansion on retrieval performance. In: Proc. of ACM SIGIR '98. (1998) 130–137
6. Conover, W.: Practical Nonparametric Statistics. 2ed edn. John Wiley & Sons, Inc (1980)
7. Davis, C.: Statistical Methods for Analysis of Repeated Measurements. Springer-Verlag NY, Inc (2002)
8. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. **22**(2) (2004) 179–214
9. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proc. of ACM SIGIR '98. (1998) 307–314