PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 103 APRIL 2015 131–138

Exact Expected Average Precision of the Random Baseline for System Evaluation

Yves Bestgen

CECL, Université catholique de Louvain

Abstract

Average precision (AP) is one of the most widely used metrics in information retrieval and natural language processing research. It is usually thought that the expected AP of a system that ranks documents randomly is equal to the proportion of relevant documents in the collection. This paper shows that this value is only approximate, and provides a procedure for efficiently computing the exact value. An analysis of the difference between the approximate and the exact value shows that the discrepancy is large when the collection contains few documents, but becomes very small when it contains at least 600 documents.

1. Introduction

Many tasks in information retrieval, computational linguistics and machine learning aim at finding relevant items among a collection of items, such as documents matching a query, subjective statements, collocations, semantic neighbors, sentences between which textual entailment holds, and so forth. To evaluate the proposed systems, precision (the proportion of retrieved documents that are relevant) and recall (the proportion of relevant documents that have been retrieved) are favored. When the system ranks the documents according to their estimated relevance, performance is typically assessed through a precision-recall curve that is summarized by average precision (AP) (Büttcher et al., 2010; Robertson, 2008). AP is equal to the average of the precision value obtained after each relevant document is retrieved (i.e., when recall increases) and corresponds to the area under the uninterpolated precision-recall

 Corresponding author: yves.bestgen@uclouvain.be
Cite as: Yves Bestgen. Exact Expected Average Precision of the Random Baseline for System Evaluation. The Prague Bulletin of Mathematical Linguistics No. 103, 2015, pp. 131–138. doi: 10.1515/pralin-2015-0007.

curve (PR) (Voorhees and Harman, 1999). More formally,

$$AP = \frac{\sum_{i=1}^{N} \frac{i}{n}}{R} \quad (1)$$

where R is number of relevant documents and n the rank of the ith document according to the system. This rank goes from 1 to N, the number of documents in the collection (see Robertson (2008, p. 689) for another, yet equivalent, formula of AP).

If AP is often used to compare the performance of different systems on the same test collection, with each system serving as benchmark for the others, the AP obtained by a system is sometimes compared to a *random baseline* AP: the expected AP that would be obtained by a system that ranks the documents in a completely random way (e.g. Marszalek et al., 2009; Nakano et al., 2011; Pecina, 2010; Pohlmeyer et al., 2011; Ramisch et al., 2008; Rasiwasia et al., 2010). This baseline AP is considered to be equal to the proportion of relevant documents in the collection, also called the category prevalence. This paper shows that this value is only approximate (section 2), and provides a procedure for efficiently computing the exact value (section 3). An analysis of the difference between the approximate and the exact value shows that the discrepancy is large when the collection contains few documents, but becomes very small when it contains at least 600 documents (section 4).

2. The Proportion of Relevant Documents is not Equal to the Expected AP for the Random Baseline

Researchers employing the proportion of relevant documents as the expected value of the random baseline AP do not justify the choice of this value. Presumably, they start from the fact that AP is equal to the area under the PR curve, and that a system that ranks the documents in a completely random way should uniformly distribute the relevant documents along the ranking. The proportion of relevant documents retrieved relative to the total number of documents considered should thus be constant at all ranking positions. It follows that the corresponding PR curve is a straight line whose intercept is the proportion of relevant documents in the collection ($p = R/N$) and whose slope is 0. The area under this "curve" is the proportion of relevant documents.

A very simple example is sufficient to show that the proportion of relevant documents is only an approximation of the actual AP for the random baseline. Consider a test collection consisting of five documents, of which two are relevant: p is thus 0.40. It is very easy to list all the possible rankings and to compute their AP as shown in Table 1. In this table, document relevance is represented by a binary variable set to one when the document is relevant. Since all these rankings are equally probable for a system that ranks documents randomly, the expected AP of the random baseline is the mean AP computed on all possible permutations. For this example, it is thus not 0.40 (R/N), but 0.593.

Ranking					AP
1st	2nd	3rd	4th	5th	
1	1	0	0	0	(1/1 + 2/2) / 2 = 1.00
1	0	1	0	0	(1/1 + 2/3) / 2 = 0.83
1	0	0	1	0	(1/1 + 2/4) / 2 = 0.75
1	0	0	0	1	(1/1 + 2/5) / 2 = 0.70
0	1	1	0	0	(1/2 + 2/3) / 2 = 0.58
0	1	0	1	0	(1/2 + 2/4) / 2 = 0.50
0	1	0	0	1	(1/2 + 2/5) / 2 = 0.45
0	0	1	1	0	(1/3 + 2/4) / 2 = 0.42
0	0	1	0	1	(1/3 + 2/5) / 2 = 0.37
0	0	0	1	1	(1/4 + 2/5) / 2 = 0.33
					Sum = 5.93
					Expected AP = 5.93 / 10 = 0.593

Table 1. Expected AP of the random baseline for $N = 5$ *and* $R = 2$

3. An Accurate and Efficient Procedure for Calculating the AP for the Random Baseline

To compute the expected AP of the random baseline for other values of N and R (or p), one could imagine using the procedure outlined in Table 1. The problem is that it would require enumerating a very large number of permutations when N is large and R is not too close to 0 or to N. It corresponds to the number of different permutations of N objects when some of these are identical, that is, $N!/(R! \times (N - R)!)$. This results in more than 17,000 billion different rankings to list for $N = 100$ and $p = 0.10$.

Looking at this table, a much more efficient solution can be proposed. If the two divisors (R and the total number of different permutations) are set aside, there remains a sum of precision scores at rank n (i.e., i/n), n corresponding to the possible positions in the ranking of each ith relevant document. For each value of i (i ranging from 1 to R), there are $N - R + 1$ possible ranks, since the ith relevant document cannot occur before the ith rank or after the $N-R+i$ rank; otherwise, there are not enough positions available for the $R - i$ remaining relevant documents. Furthermore, for each value of i, there are in theory a total of $N!/(R! \times (N - R)!)$ precision scores to compute, since the ith relevant document is present in every possible permutation. But the problem can be reformulated in terms of the probability that the ith relevant document occurs at each of the $N - R + 1$ possible ranks, all the probabilities computed for a given i summing to 1. This formulation requires the calculation of only $R \times (R+N-1)$ values.

i	n	dhyper(i,N,N-R,n)	i/n	Final prob.	P@n	Contribution to AP
1	1	0.4	1.00	0.4	1.00	0.400
1	2	0.6	0.50	0.3	0.50	0.150
1	3	0.6	0.33	0.2	0.33	0.067
1	4	0.4	0.25	0.1	0.25	0.025
2	2	0.1	1.00	0.1	1.00	0.100
2	3	0.3	0.67	0.2	0.67	0.134
2	4	0.6	0.50	0.3	0.50	0.150
2	5	1.0	0.40	0.4	0.40	0.160
						Sum = 1.186
						Expected AP = 1.186 / 2 = 0.593

Note: *dhyper()* returns the density for the hypergeometric function. *P@n* stands for *precision at rank n*.

Table 2. Calculation of the expected AP of the random baseline for $N = 5$ *and* $R = 2$

The proposed procedure is, therefore, to calculate, for each rank n, the probability that the ith relevant document occurs at that rank, producing a precision score at rank n equal to i/n. This probability is equal to the probability of having i successes in n draws *without replacement* from a population of size N containing R successes and $N - R$ failures, with the additional condition that the ith success occurs at the last draw (i.e., at rank n). The first part of this probability is given by the hypergeometric distribution whose formula is:

$$P(X = i) = \frac{\binom{R}{i}\binom{N-R}{n-i}}{\binom{N}{n}} \tag{2}$$

where $\binom{R}{i}$ is a binomial coefficient, corresponding here to $i!/(R!(R-i)!)$. Regarding the additional condition, the probability of a success at the last draw when there are i successes in n draws is obviously i/n. Multiplying this final probability by the precision score at rank n produces the contribution of each ith relevant document to the total sum of AP, and it only remains to divide this sum by R to obtain the expected AP of the random baseline.

Table 2 applies this calculation procedure[1] to the example of Table 1. In this example, the gain in number of operations is very small (eight instead of 10), but for $N = 100$ and $R = 10$, it is reduced from more than 17,000 billion to 910. Calculating

[1] In this table, the *Final prob.* values for each i sum to 1, as explained. This is not the case for the probabilities from the hypergeometric distribution alone, which sum to 1 for a given n only if one adds the probabilities for all possible i (i.e., the number of successes) including 0 success.

a hypergeometric probability takes more time than finding a possible permutation, but extremely efficient procedures for calculating these probabilities are available in every major statistical software.

The very simple R function (R Core Team, 2013) given below implements the calculation procedure of the expected AP of the random baseline for any values of N and R.

```
RandomAPExact = function(N=0, R=0) {
  ap = 0
  for (i in 1:R) {
   for (n in i:(N-R+i)) {
    ap = ap + dhyper(i,R,N-R,n)*(i/n)*(i/n)
   }
  }
  ap = ap/R
  ap
 }
Function call: RandomAPExact(10,4)
Result: [1] 0.5285979
```

For N = 100 and p = 0.10, this function takes 0.011 seconds to compute the solution, and just over 131 seconds for N = 10000 and p = 0.40 on an *Intel Core i5* 2.66GHz processor. If there is no doubt that the exact procedure is computationally intensive compared to the calculation of the approximate value, the R code provided allows to calculate it very easily and it should only be used once for an evaluation task. What is a handful of minutes compared to the time required for the development of an IR system and for its evaluation?

4. How Large is the Difference Between the Exact Value and the Usual Estimate?

To get an idea of the importance of differences between the exact AP and the approximate AP for the random baseline, Figure 1 shows the evolution of this difference for many values of N and p, the exact value being systematically larger than the approximate value. As can be seen, the difference decreases when N increases or p increases. As soon as N is at least equal to 600, it is less than 0.01 for all tested values of p.

5. Conclusion

This paper shows that the AP for the random baseline usually used in information retrieval and computational linguistics is only an approximation of the exact AP, and it presents an efficient procedure to compute the latter. An analysis of the dif-

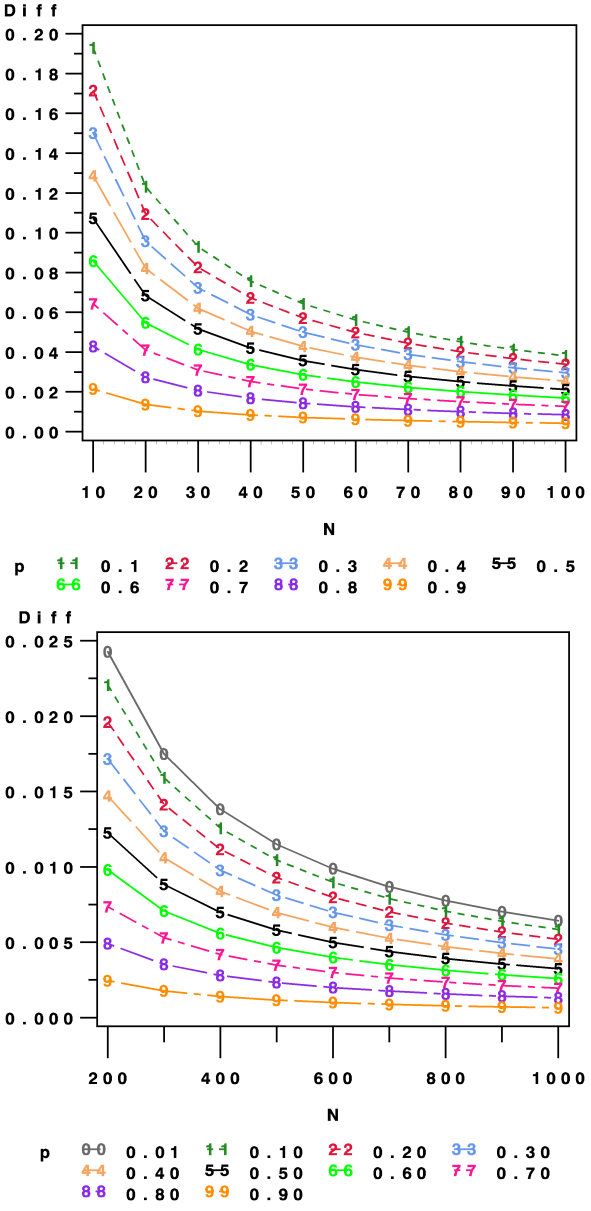

Figure 1. Difference between the exact and the approximate AP of the random baseline for several values of N *and* p

ference between the exact value and the approximate value shows that the discrepancy between them reduces when the size of the collection of documents increases. While many evaluations of IR systems are performed on very large collections of documents, some research areas use much smaller collections because of the difficulties encountered in their constitution (i.e., less resourced languages, emerging tasks or tasks requiring complex relevance judgment that can only be performed by human experts). The smallness of the collections can be further enhanced by the use of a random under-sampling procedure advocated by Jeni et al. (2013) to reduce the impact of large imbalance between the positive and negative examples on performance metrics.

In conclusion, it can be recommended that researchers who plan to compare their system to the random baseline AP use the proposed procedure to calculate the exact expected value when the test collection is of limited size or, when there are at least 600 documents in the collection, state in their report that the proportion of relevant documents in the collection is an excellent approximation of the exact value.

Acknowledgements

Yves Bestgen is a Research Associate with the Belgian Fund for Scientific Research (F.R.S-FNRS).

Bibliography

Büttcher, Stefan, Charles Clarke, and Gordon Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010.

Jeni, László, Jeffrey Cohn, and Fernando De La Torre. Facing imbalanced data - recommendations for the use of performance metrics. In *ACII '13: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251, 2013.

Marszalek, Marcin, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936, 2009.

Nakano, Takuho, Akisato Kimura, Hirokazu Kameoka, Shigeki Miyabe, Shigeki Sagayama, Nobutaka Ono, Kunio Kashino, and Takuya Nishimoto. Automatic video annotation via hierarchical topic trajectory model considering cross-modal correlations. In *Acoustics, Speech and Signal Processing*, pages 2380–2383, 2011.

Pecina, Pavel. Lexical association measures and collocation extraction. *Language Resources & Evaluation*, 44:137–158, 2010.

Pohlmeyer, Eric, Jun Wang, David Jangraw, Bin Lou, Shih-Fu Chang, and Paul Sajda. Closing the loop in cortically-coupled computer vision: a brain–computer interface for searching image databases. *Journal of Neural Engineering*, 8:1–14, 2011.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2013.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. An evaluation of methods for the extraction of multiword expressions. In *LREC Workshop towards a Shared Task for Multiword Expressions*, pages 50–53, 2008.

Rasiwasia, Nikhil, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM '10, the International Conference on Multimedia*, pages 251–260, 2010.

Robertson, Stephen. A new interpretation of average precision. In *31st Annual international ACM SIGIR Conference*, pages 689–690, 2008.

Voorhees, Ellen and Donna Harman. Overview of the seventh text retrieval conference. In *Seventh Text REtrieval Conference (TREC-7)*. NIST Special Publication, 1999.

Address for correspondence:
Yves Bestgen
yves.bestgen@uclouvain.be
Centre for English Corpus Linguistics
Université catholique de Louvain
Place du cardinal Mercier, 10
Louvain-la-Neuve, 1348, Belgium