Precision and Recall Without Ground Truth

Bart Lamiroy, Tao Sun

Precision and Recall Without Ground Truth

Bart Lamiroy*
Nancy Université – INPL – LORIA
Nancy, France
Bart.Lamiroy@loria.fr

Tao Sun
Lehigh University
Computer Science and Engineering
Bethlehem, PA

June 7, 2011

Abstract

In this paper we present a way to use precision and recall measures in total absence of ground truth.

1 Precision and Recall

1.1 General Definitions and Notation

Precision Pr and Recall Rc (and often associated F-measure or ROC curves) are standard metrics expressing the *quality* of Information Retrieval methods [8]. They are usually expressed with respect to a query q (or averaged over a series of queries) over a data set Δ such that:

$$Pr_q^\Delta = \frac{\left|\mathcal{P}_q^\Delta \cap \mathcal{R}_q^\Delta\right|}{\left|\mathcal{R}_q^\Delta\right|} \tag{1}$$

$$Rc_q^\Delta = \frac{\left|\mathcal{P}_q^\Delta \cap \mathcal{R}_q^\Delta\right|}{\left|\mathcal{P}_q^\Delta\right|} \tag{2}$$

where $\mathcal{P}_q^\Delta$ is the set of all documents in Δ, relevant to query q, and where $\mathcal{R}_q^\Delta$ is the set of documents actually retrieved by q. Although we can make a safe assumption by considering $\mathcal{R}_q^\Delta$ known (*i.e* the query q can actually be executed, and returns a known, manageable set of results), the same assumption does not always hold for $\mathcal{P}_q^\Delta$, as will be shown later. For ease of reading we will refer to respectively Pr, $\mathcal{P}$ Rc, and $\mathcal{R}$, when there is no ambiguity on Δ and q.

Often both are combined in the F_β measure, where

$$F_\beta = \left(1+\beta^2\right)\frac{PrRc}{\beta^2 Pr + Rc} \tag{3}$$

*Bart Lamiroy was a visiting scientist at Lehigh University in 2010-2011. This work was conducted at the Computer Science and Engineering Departent at Lehigh University and was supported in part by a DARPA IPTO grant administered by Raytheon BBN Technologies.

1.2 Other Interpretations and Frameworks

Precision, Recall and the F-measure can also be defined with respect to *true positives* τ_p, *false positives* ϕ_p, *true negatives* τ_n and *false negatives* ϕ_p. In that case, the corresponding formulas are:

$$Pr = \frac{\tau_p}{\tau_p + \phi_p} \tag{4}$$

$$Rc = \frac{\tau_p}{\tau_p + \phi_n} \tag{5}$$

$$F_\beta = \frac{\left(1 + \beta^2\right) \tau_p}{\left(1 + \beta^2\right) \tau_p + \beta^2 \phi_n + \phi_p} \tag{6}$$

Here again, it is necessary to know the values of τ_p, ϕ_p, τ_n and ϕ_p (as, previously, the sets $\mathcal{P}$ and $\mathcal{R}$) in order to be able to do the computations.

It is also possible to give probabilistic interpretations to Pr and Rc. In that case, Pr would be the probability that a random document retrieved by the query is relevant, and Rc that a random relevant document be retreived by the query (taking as assumption that documents have uniform distributions). This is the interpretation we are going to use in the next sections.

2 Absence of Ground Truth

Previously enumerated metrics all made the assumption that the returns of queries can, in some way be qualified as "good" or "bad". Most often, there even is the assumption that this can actually be quantified: belonging to set $\mathcal{P}$, τ_p, *etc.* This implies that there is some absolute knowledge of *ground truth* or an *oracle* function available for the assessment of these quantities. While it is very convenient to rely on established truth to further train or evaluate methods, it is often very costly to obtain in many cases, and even impossible in others. Furthermore, it generally requires some human intervention or validation of some sorts, which makes the ground-truthing process both difficultly scalable and error prone, and therefore costly.

This paper presents a way to estimate precision and recall using a probabilistic model, allowing either to compare algorithms operating on the same data, without the requirement of establishing ground truth, or, to leverage crowdsourcing to establish ground truth in presence of noise, errors and mistakes. In order to achieve this, we shall first establish the underlying assumptions to our approach, in section 2.1, defining the context in which we have conceived our model. We then develop the mathematical foundations an tools in section 2.2.1.

2.1 General Assumptions

In what follows we are assuming that the following general conditions and notations apply:

1. We are considering generic system $\mathcal{S}$ that, given a query q, partitions[1] a set of documents $\Delta = \{\delta_i\}_{i=1\ldots d}$ into $\mathcal{S}^{q+}$ and $\mathcal{S}^{q-}$.

[1] For the absent-minded reader, "*partitioning*" Δ into $\mathcal{S}^+$ and $\mathcal{S}^-$ entails that $\Delta = \mathcal{S}^+ \cup \mathcal{S}^-$ and $\mathcal{S}^+ \cap \mathcal{S}^- = \emptyset$

The partitioning function $\mathcal{S}^q$ is defined as

$$\begin{aligned} \mathcal{S}^q : \Delta &\to \{+, -\} \\ \delta_i &\mapsto \mathcal{S}^q(\delta_i) \end{aligned} \tag{7}$$

$\mathcal{S}^{q+}$ (resp. $\mathcal{S}^{q-}$) is defined as the inverse image of $\{+\}$ (resp. $\{-\}$).

2. Other systems, similar to $\mathcal{S}^q$ exist and their partitioning results are available. It is assumed that these systems operate in the same semantic context, and therefore aim to achieve the same partitioning as $\mathcal{S}^q$. We shall refer to the set of these systems as $\Sigma^q = \{\mathcal{S}_i^q\}_{i=1\ldots s}$

In what follows, and where it is obvious, parameter q will be omitted. Table 1 gives an example overview of what three different systems could produce for a given query over a particular document set Δ.

Δ	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$
δ_1	+	+	+
δ_2	+	+	+
δ_3	-	+	-
δ_4	+	-	-
δ_5	+	-	-
δ_6	-	-	+
δ_7	-	-	-

$\mathcal{S}_1^+ = \{\delta_1, \delta_2, \delta_4, \delta_5\}$
$\mathcal{S}_1^- = \{\delta_3, \delta_6, \delta_7\}$

$\mathcal{S}_2^+ = \{\delta_1, \delta_2, \delta_3\}$
$\mathcal{S}_2^- = \{\delta_4, \delta_5, \delta_6, \delta_7\}$

$\mathcal{S}_3^+ = \{\delta_1, \delta_2, \delta_6\}$
$\mathcal{S}_3^- = \{\delta_3, \delta_4, \delta_5, \delta_7\}$

Table 1: Example of query systems $\mathcal{S}_i$ operating on document set Δ

2.2 Performance Evaluation

The question that arises now is how to compare different $\mathcal{S}_i$ and decide which one performs best. Traditionally, one would take an evaluation test set $\Delta_\star$ for which the ground truth of a query $q_\star$ is known and available. We shall refer to this ground truth as $\Delta_\star^+$ and $\Delta_\star^-$ (*i.e.* $\Delta_\star^+$ is the partition of $\Delta_\star$ containing the documents corresponding to $q_\star$, $\Delta_\star^-$ its complement). This knowledge then allows to compute precision and recall values, as described in Section 1, for all $\mathcal{S}_i$ and establish a performance metric adapted to the context under consideration.

When $\Delta_\star^+$ and $\Delta_\star^-$ are unavailable, it is less obvious to compare the results of the different $\mathcal{S}_i$. One well documented approach is to use statistical estimators by considering each $\mathcal{S}_i(\Delta)$ as the outcome of some random variable. What we are going to develop here, is very similar, but particularly focused on the expression of precision and recall.

2.2.1 Simplified Case

First we're making the assumption that all $\mathcal{S}_i$ are of equal importance, and that there is no *a priori* knowledge available allowing to presume some of the systems are more reliable than others. This assumption will be alleviated in later work. We also assume all documents have equal frequency and occurrence probability.

For the arguments developed next, we need to introduce two "virtual" query systems, $\mathcal{S}_\top$ and $\mathcal{S}_\bot$. $\mathcal{S}_\top$ always returns all documents for any given query, $\mathcal{S}_\bot$ never returns any. In other terms,

$$\mathcal{S}_\top^+ = \Delta, \mathcal{S}_\top^- = \emptyset \tag{8}$$

$$\mathcal{S}_\bot^+ = \emptyset, \mathcal{S}_\bot^- = \Delta \tag{9}$$

We are also slightly reconsidering the partitioning function defined in equation (7), such that it returns values in $\{1, 0\}$ rather than in $\{+, -\}$.

Under these hypotheses, the probability that a document δ_i belongs to $\Delta_\star^+$ is

$$P(\delta_i) = \frac{1}{s+2} \sum_{k=1\ldots s,\bot,\top} S_k(\delta_i) \tag{10}$$

The results of the application of this to the example in Table 1, is represented in Table 2.

Δ	$P(\delta_i)$	$\mathcal{S}_\top$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_\bot$
δ_1	0.8	1	1	1	1	0
δ_2	0.8	1	1	1	1	0
δ_3	0.4	1	0	1	0	0
δ_4	0.4	1	1	0	0	0
δ_5	0.4	1	1	0	0	0
δ_6	0.4	1	0	0	1	0
δ_7	0.2	1	0	0	0	0

Table 2: Example

Given the hypothesis of equidestribution of all documents δ_i in Δ and given the probabilistic definition of precision in Section 1.2, stating that Pr "is the probability that a random document retrieved by a query is relevant", we can now define $Pr(\mathcal{S}_k)$:

$$Pr(\mathcal{S}_k) = \frac{\sum_{i=1\ldots d} P(\delta_i) S_k(\delta_i)}{\sum_{i=1\ldots d} S_k(\delta_i)} \tag{11}$$

Similarly, Rc was defined as "the probability for a random relevant document to be retrieved by the query". In our case, however relevancy has no longer a binary value, but has been replaced by $P(\delta_i)$. By reformulating this conditional probability and using Bayes' theorem (and using the fact that the

inverse conditional of Rc is Pr), things smooth out elegantly.

$$\begin{aligned}
Rc(\mathcal{S}_k) &= Prob\left(\text{retrievedBy}_{\mathcal{S}_k}(\delta_i) \,\middle|\, \text{isRelevant }_{(\delta_i)}\right) \\
&= Prob\left(\text{isRelevant}(\delta_i) \,\middle|\, \text{retrievedBy}_{\mathcal{S}_k(\delta_i)}\right) \frac{Prob\left(\text{retrievedBy}_{\mathcal{S}_k}(\delta_i)\right)}{Prob\left(\text{isRelevant}(\delta_i)\right)} \\
&= Pr(\mathcal{S}_k) \frac{\frac{1}{d}\sum_{i=1..d} S_k(\delta_i)}{\frac{1}{d}\sum_{i=1..d} P(\delta_i)} \\
&= \frac{\sum_{i=1...d} P(\delta_i) S_k(\delta_i)}{\sum_{i=1...d} S_k(\delta_i)} \frac{\sum_{i=1..d} S_k(\delta_i)}{\sum_{i=1..d} P(\delta_i)} \\
&= \frac{\sum_{i=1...d} P(\delta_i) S_k(\delta_i)}{\sum_{i=1..d} P(\delta_i)} \qquad (12)
\end{aligned}$$

It is interesting to notice the resemblance between equations (1) and (11) as well as between (2) and (12). Table 3 shows the values obtained when applied to the examples of Table 2.

Δ	$P(\delta_i)$	$\mathcal{S}_\top$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_\bot$
δ_1	0.8	1	1	1	1	0
δ_2	0.8	1	1	1	1	0
δ_3	0.4	1	0	1	0	0
δ_4	0.4	1	1	0	0	0
δ_5	0.4	1	1	0	0	0
δ_6	0.4	1	0	0	1	0
δ_7	0.2	1	0	0	0	0
Sum	3.4	7	4	3	3	0
$\sum P\mathcal{S}_k$		3.4	2.4	2	2	0
Pr		0.49	0.6	0.67	**0.67**	∞
Rc		1	**0.71**	0.59	0.59	0

Table 3: Example of precision and recall computations without established ground truth.

2.3 Experimental Validation

In order to experimentally validate the model developed we have taken two contexts. One consists in taking the results of experiments reported in [4] related to comparing standard symbol recognition techniques. A second is related to evaluation of binarization algorithms on downstream treatment.

2.3.1 Symbol Recognition

In this section we use the experimental results reported in [4]. In this paper, the authors compare 5 different symbol recognition methods on a set of electrical wiring diagrams. Since their dataset has no know ground truth, they use a panel of human annotators to select and determine which ground truth corresponds to which query.

k	Radon	GFD	zernike	SC	ARG
1	1.00000	1.00000	1.00000	1.00000	1.00000
2	0.59524	0.93023	0.72727	0.90698	0.97674
3	0.43548	0.80000	0.60000	0.82813	0.92188
4	0.34940	0.70930	0.49412	0.75581	0.83333
5	0.30769	0.66038	0.44762	0.69811	0.81905
6	0.28226	0.58730	0.41600	0.62698	0.75397
7	0.26712	0.53061	0.37838	0.56164	0.72789
8	0.23952	0.49102	0.35714	0.51205	0.68452
9	0.21925	0.45455	0.35106	0.47059	0.65263
10	0.20290	0.41546	0.33173	0.43269	0.61321

Table 4: Precision measures as reported in [4].

k	Radon	GFD	zernike	SC	ARG
1	0.05699	0.05699	0.05699	0.05699	0.05699
2	0.06476	0.10362	0.08290	0.10103	0.10880
3	0.06994	0.13471	0.10103	0.13730	0.15284
4	0.07512	0.15803	0.10880	0.16839	0.18134
5	0.08290	0.18134	0.12176	0.19170	0.22279
6	0.09067	0.19170	0.13471	0.20466	0.24611
7	0.10103	0.20207	0.14507	0.21243	0.27720
8	0.10362	0.21243	0.15544	0.22020	0.29792
9	0.10621	0.22020	0.17098	0.22797	0.32124
10	0.10880	0.22279	0.17875	0.23316	0.33678

Table 5: Recall measures as reported in [4].

Since the authors in [4] report retrieval efficiency, as defined in [3], we have resampled their raw experimental data to extract precision and recall.

The results, with respect to the human-defined ground-truth reported by the authors is given in Tables 4 and 5. These data are also presented graphically in Figure 1.

Figure 2 reproduces the precision and recall values obtained using our method on the exact same data. It is interesting to note that, with one noteworthy exception, the ordering of the tested methods, with respect to precision or recall (*i.e.* when ordering methods from high precision/recall to low) is respected. Although not reproduced here, this also holds for the F-measure. What is even more compelling, is that the methods 'SC' and 'GFD' maintain their similarity in both cases, with and without consideration of ground truth.

The one exception is the 'ARG' method. While considered as a tie with 'SC' and 'GFD' with our method, it significantly outperforms all other approaches according to the ground truth. This is a very interesting result, and is currently under investigation.

2.3.2 Document Binarization

The data used in this second study are the historical images collected from the Library of Congress on-line data set[1]. A total of 60 TIF format images

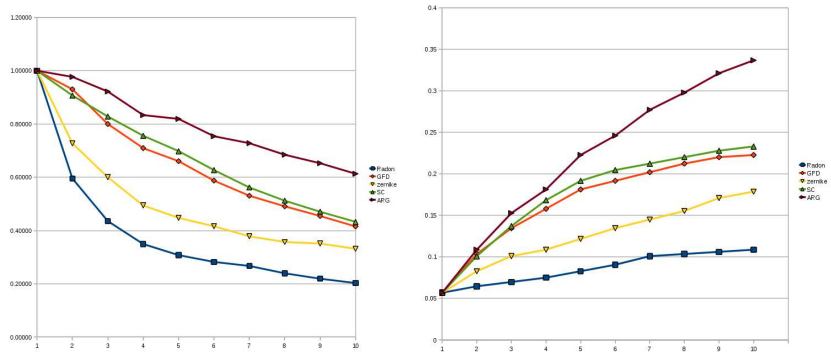

Figure 1: Precision and Recall as reported in [4]

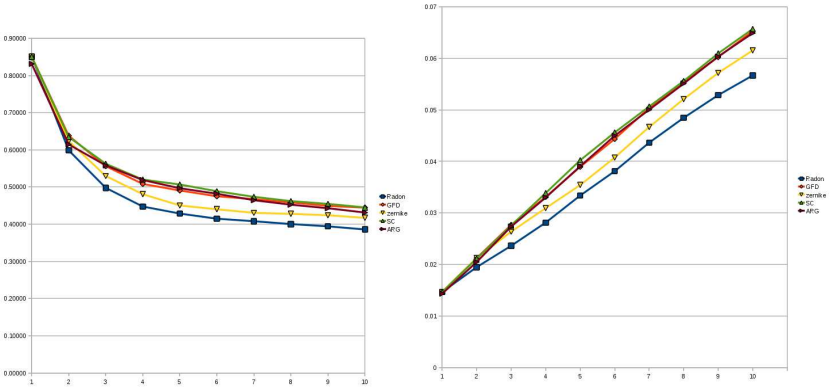

Figure 2: Precision and Recall as computed without ground truth

with a resolution of 300 dpi. Various genres from official documents to private letters are included. The degraded quality of these images, such as uneven illumination, bleeding-through, handwritten marks, *etc*, are be a great challenge for recognition algorithms. In this case, we are going to try and use our approach to evaluating binarization quality to downstream recognition. The document image analysis pipeline consists of three stages:

1. Binarization;
2. OCR;
3. Named Entity Recognition

	Otsu	Sauvola	Wolf
Precision	0.6223	0.7715	0.7533
Recall	0.5915	0.7281	0.7230

Table 6: Average Recognition Accuracies with Ground Truth

Table 7: Method I: Average Recognition Accuracies without Ground Truth

	$S^{\top}$	Otsu	Sauvola	Wolf	$S_{\perp}$
Precision	0.4000	0.6327	0.6757	0.6722	∞
Recall	1.0000	0.5153	0.5660	0.5662	0

Table 8: Method II: Average Recognition Accuracies without Ground Truth

	$S^{\top}$	Otsu	Sauvola	Wolf	$S_{\perp}$
Precision	0.5733	0.6035	0.6450	0.6416	∞
Recall	1.0000	0.6550	0.6988	0.6957	0

Binarization is the first stage, and three thresholding methods are used in this stage respectively. They are Otsu[5], Sauvola[6] and Wolf[9]. Otsu's method is a global thresholding method while the latter two are local thresholding methods. After all the images are converted into binary images, the resultant binary images were converted to ASCII texts by Tesseract-3.00[7] open source software package in the second stage. Finally, Stanford Named Entity Recognizer [2] is used in the third stage. To sum up, we have three different pipelines this way. Although our method aims to calculate precision and recall without ground truth, we still need ground truth to evaluate if our method can achieve the goal proposed in Section 2.2. Since the groundtruth of the historical images are not directly available, we generate the groundtruth ourselves by manual typing the text and carefully proofreading.

Since the three different pipelines depend on three different thresholding methods, we use the names of them to stand for the three pipelines, respectively. The calculation of average precision and recall is based on the outputs of these pipelines, which are the named entity extraction results. When evaluating our method, we use two different ways to process the outputs of the three pipelines. Method I considers all the recognized named entities as "bag-of-words", so they are organized in an alphabetical way. While Method II use the multiple sequence alignment algorithm to align the three outputs first, the original positions of these named entities are kept this way. The experiment results are shown in the following tables. From Table 6 we can see that Sauvola and Wolf beat Otsu thresholding method. The reason is obvious. Only one threshold is determined for the whole image by Otsu, while for the other two methods, different thresholds are calculated according to the grey distribution of their corresponding local windows. Table 7 and Table 8 show the results of our method under two methods. We can see again the performance of Sauvola and Wolf is better than that of Otsu, while recognition accuracies between Sauvola and Wolf are similar. Both of them indicate that even if without ground truth, the precision and recall computed by our method is similar to those computed with ground truth.

3 Conclusion and Future Work

In this study we have presented how to compute precision and recall without presence of formally identified ground truth. Results indicate that this measure is coherent with real, ground truth based precision and recall measures.

Further work and development will consist in establishing:

1. how to rank or take into account user-contributed "partial" ground truth, especially considering "yes/no/unknown" information ?
2. What if we use confidence measures instead of binary values on the algorithm outcomes ?
3. What if we attach more weight or confidence to particular algorithms ?

Acknowledgements

The authors would like to acknowledge Santosh K.C. for having provided the experimental data, used in Section 2.3.1.

References

[1] Library of congress. http://memory.loc.gov/.

[2] Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*. The Association for Computer Linguistics, 2005.

[3] Mohan S. Kankanhalli, Babu M. Mehtre, and Jian Kang Wu. Cluster-based color matching for image retrieval. *Pattern Recognition*, 29:701–708, 1995.

[4] Santosh K.C., Bart Lamiroy, and Laurent Wendling. Spatio-structural symbol description with statistical feature add-on. In *The Ninth International Workshop on Graphics Recognition (GREC2011)*, 2011. submitted.

[5] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, January 1979.

[6] Jaakko J. Sauvola and Matti Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.

[7] Ray Smith. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA, 2007. IEEE Computer Society.

[8] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[9] Christian Wolf and David S. Doermann. Binarization of low quality text using a markov random field model. In *ICPR (3)*, pages 160–163, 2002.