Evaluation of Criteria for Information Retrieval

Tsunenori Ishioka
The National Center for University Entrance Examinations, Japan
tunenori@rd.dnc.ac.jp

Abstract

We investigate van Rijsbergen's F-measure, break-even point, and 11-point averaged precision, all of which can be translated into 1-dimensional scalar quantity from the precision and the recall. These investigations can be done by comparing to tetrachoric (four-fold) correlation coefficient and phi (four-fold point) coefficient, which are often used as the index of statistical association in a 2×2 contingency table. The results show that when a fallout rate is less than 0.1, (1) the F_1 measure has similar properties of the phi coefficient, (2) the break-even point is almost equivalent to a phi coefficient, and (3) the 11-point averaged precision should be a measure which is larger than a phi coefficient and has a value smaller than a tetrachoric correlation coefficient.

1. Introduction

Due to the increased importance of the Internet, the use of the search engines like Yahoo (http://www.yahoo.com) and Google (http://www.google.com) is becoming increasingly widespread among all office workers. Consequently, the reliability of the information retrieval results should be discussed in greater detail.

In information retrieval, the first of two primary evaluation measures is recall, which shows the ability of a retrieval system to present all relevant items, and the second is precision, which shows the ability of a retrieval system to present only relevant items. As a comprehensive measure of recall and precision, the F-measure of van Rijsbergen is widely used. Moreover, the break-even point and 11-point average precision, which are mentioned later, are measures that are also used for mutual comparison of search engines and retrieval systems. These indices are considered the de-facto standards for evaluation criteria in information retrieval [12].

However, no papers about the statistical properties of the evaluation criteria themselves are included in the collection of papers presented over the last ten years of the most prominent international meeting about document retrieval, the Text REtrieval Conference (TREC). Furthermore, there are few related references in referred papers at the international meetings of SIGIRs on information retrieval between 1998 to 2001 as sponsored by the Association for Computing Machinery (ACM). (The only two references to evaluation criteria were made by Fujita[2], which mentioned the 'aboutness' of Ingwersen[4], and Goldstein[3], which proposed evaluation criteria for document summarizing.)

In this paper, we explore the relationship between the evaluation measures used for information retrieval systems and the evaluation measures of statistical 2×2 contingency table.

Section 2 summarizes the evaluation measure used by information retrieval systems. Section 3 explains the statistical basis of the evaluation measure used in the 2×2 contingency table. This section also describes the comparative examination with different evaluation measure used by information retrieval system. Section 4 concludes the paper.

2 Evaluation measures in information retrieval

2.1 Recall and Precision

Suppose that the relevance of a document to a retrieval query could be given regarding a certain document set. Were this possible, a cross matrix like Table 1 could be prepared for the retrieval query. The matrix shows whether a row matches with a retrieval query, and whether the column was searched by the retrieval query. The each element of the matrix is the number of documents.

Table 1. Cross matrix

	retrieved	not retrieved	total
relevant	f_{11}	f_{12}	$f_{1\cdot}$
nonrelevant	f_{21}	f_{22}	$f_{2\cdot}$
total	$f_{\cdot1}$	$f_{\cdot2}$	$f_{\cdot\cdot} = n$

When Table 1 is given, recall r and precision p are defined as follows:

$$\text{recall}: \quad r = \frac{f_{11}}{f_{1\cdot}}, \tag{1}$$

$$\text{precision}: \quad p = \frac{f_{11}}{f_{\cdot 1}}. \tag{2}$$

That is, a recall is the percentage of relevant documents retrieved[5], and shows a "leak" in retrieval. On the other hand, precision is the percentage of retrieved documents that are relevant[5], and shows a "noise" in retrieval.

Therefore, although it is better for the recall and precision values to be large, there is an inverse relationship between them. If the recall is going to be high, the precision will be low. The opposite is also true.

The index of fallout may be used instead of a recall. Fallout is defined by the following formula:

$$\text{fallout}: \ a = \frac{f_{21}}{f_{2\cdot}}. \tag{3}$$

Fallout is the percentage of nonrelevant documents that were retrieved[5]. Fallout expresses the error of a system and can be called the user-oriented measure; while a recall shows whether no leak occurred while the relevant documents were searched, and can be considered to be the system-oriented measure.

Reference is made to the practical range of values of recall, precision, and fallout. According to Belkin[1], recall and precision in an actual system are 60 % and 40 %, at best. In fact, when seeing the performance of Web retrieval of the latest TREC-9[11], we find completely automatic at the best system (ric9dpn), whose precision is 27% to recall 40 %; a user's feedback can be put in, and precision can be raised to 60 %. If the term "information retrieval" is searched with the newest Web retrieval experiment service http://infobee.ne.jp/ currently offered by NTT, 117,240 sites can be searched out of the total number of Web pages now numbering 40 million.

When recall is 40% and precision is 27%, the elements of the cross matrix are $f_{11} = 117\ 240,\ f_{12} = 175\ 860,\ f_{21} = 316\ 982,\ f_{22} = 39\ 389\ 918$ and fallout a is

$$a = 316\ 982/39\ 706\ 900 = 0.008.$$

When the term " information retrieval & reliability" is specified, 1,364 sites are hit. The resultant fallout is

$$a = 3\ 688/39\ 996\ 590 = 0.000\ 09.$$

In information retrieval, the number of nonrelevant documents usually outnumbers that of relevant documents. This example shows that a narrowing down of $a = 0.01$ is insufficient, since fallout shows the rate searched from vast quantity of nonrelevant documents. Until now, $a < 0.01$ has been an adequate fallout value.

2.2 Summary value of recall and precision

Van Rijsbergen's F measure is known as a method of changing a recall r and precision p into 1-dimensional scalar:

$$F_\beta = \frac{(1+\beta^2)pr}{\beta^2 p + r} = \frac{(1+\beta^2)f_{11}}{\beta^2 f_{1\cdot} + f_{\cdot 1}}, \tag{4}$$

where β indicates relative importance.

For example, $\beta = 1$ represents a to the same extent precision of a recall, and $\beta = 2$ represents a precision as important as twice compared with a recall.

At the time of $\beta = 1$, especially a formula (4) is set to

$$F_1 = \frac{2pr}{p+r}, \tag{5}$$

and is in agreement with the harmonic mean of recall and precision. F measure indicates that the larger value produces better retrieval. The two evaluation measures often used by some conferences like TREC other than an F measure, are break-even point and 11-point averaged precision.

- break-even point

  The break-even point is the point at which a retrieval's recall and precision correspond., i.e., the point when the straight line of inclination 1 is crossed on a recall-precision curve. However, since an actual plotting does not result in a smooth curve, suitable interpolation is needed.

- 11-point averaged precision

  An 11-point averaged precision is completed by averaging the precision at 11 standard recall levels $(0.0, 0.1, 0.2, \ldots, 1.0)$. The precision at recall level 0.0 cannot be found theoretically, so it is approximated using the precision value at which a relevant document was first searched. The 11-point average precision is more comprehensive than the break-even point.

More details about these evaluation criteria can be referred to 'Evaluation Techniques and Measures' in an appendix of TREC-8[10].

3 Index of relevance

The 2×2 contingency table shown in Table 1 has been considered for many years to be a special case of the $k \times \ell$ contingency table. For example, 21 indices as a relevance index between binary variates are shown in 'dictionary in statistics' [9].

The most general relevance indices for binary variates are the tetrachoric (four-fold) correlation coefficient and the phi (four-fold point) coefficient. The former shows when both variables of x and y under the normal distribution are individually divided, how the correlation ρ of two variables before division had been acted. In order to find ρ, a computer program, for example 116.f [8], is required.

The phi coefficient applies the definition of Pearson's product-moment correlation coefficient, under the condition that variables x and y can take only binary values. The phi coefficient is expressed as

$$\widehat{\phi} = \frac{f_{11} f_{22} - f_{21} f_{12}}{\sqrt{f_{1\cdot} f_{2\cdot} f_{\cdot 1} f_{\cdot 2}}}. \tag{6}$$

3.1 Comparison of F measure with a tetrachoric or phi coefficient

Generality is not lost even if we fix total number of data n in the contingency table in Table 1. However, in the case of information retrieval, we need to be cautious of circumference frequency ($f_{\cdot 1}$ and $f_{\cdot 2}$) being unfixable. For example, a vector space model [7] retrieves a document whose indicating vector is similar to an inquiry vector; in fact, this model searches the document beyond a certain threshold which shows the degree to be similar. The results depends on a setting of this threshold, and each total number of retrieved or not retrieved will be varied. Also, the judgment of the relevant or nonrelevant documents cannot be absolute. Thus, a statistical analysis which, for example, assumed the hypergeometric distribution of f_{11}, is unsuitable.

If we temporarily set f_{11} as known, we obtain the following by using $f_{\cdot 1} = f_{11}/p$,

$$f_{21} = f_{\cdot 1} - f_{11} = f_{11}\left(\frac{1}{p} - 1\right). \tag{7}$$

In the same way, we get

$$f_{12} = f_{11}\left(\frac{1}{r} - 1\right). \tag{8}$$

Moreover, since $f_{2\cdot} = f_{21}/a$ is possible, we can show that

$$f_{22} = f_{2\cdot} - f_{21} = f_{11}\left(\frac{1}{p} - 1\right)\left(\frac{1}{a} - 1\right). \tag{9}$$

Since n can be shown as

$$n = f_{11} + f_{12} + f_{21} + f_{22}, \tag{10}$$

we can find f_{11} from Eq. (7) – (10), by

$$f_{11} = n/\left\{\frac{1}{p} + \frac{1}{r} - 1 + \left(\frac{1}{p} - 1\right)\left(\frac{1}{a} - 1\right)\right\}. \tag{11}$$

If we assume that f_{11} is the function of only r, f_{11} can be written as

$$f_{11} = \frac{n}{\frac{1}{r} + c}, \tag{12}$$

using a constant c. If a partial differential is carried out by r, we get

$$\frac{\partial f_{11}}{\partial r} = \frac{n}{(1 + cr)^2} > 0. \tag{13}$$

It turns out that f_{11} is a uniform increase of r. A uniform increase of p also occurs.

Recall r, precision p, and fallout a, were given in Eq. (11), (7)–(9), then $f_{11}, f_{12}, f_{21}, f_{22}$ express correctly, we can calculate a tetrachoric correlation coefficient ρ and phi coefficient ϕ. Figures 1 and 2 show the tetrachoric correlation coefficient ρ and phi coefficient ϕ. The coefficients are shown on the z axis; the precision p is shown on the x axis, and the recall r on the y axis. The label (a), (b) and (c) in the figures correspond to the fallout values of $a = 0.01, 0.1, 0.2$.

Generally, a tetrachoric correlation coefficient and a phi coefficient are similar indices. However, in the range of the usual information retrieval namely in $a \leq 0.1$, they are not. In this range, the tetrachoric correlation coefficient has a much larger value than the phi coefficient, as shown in the figures.

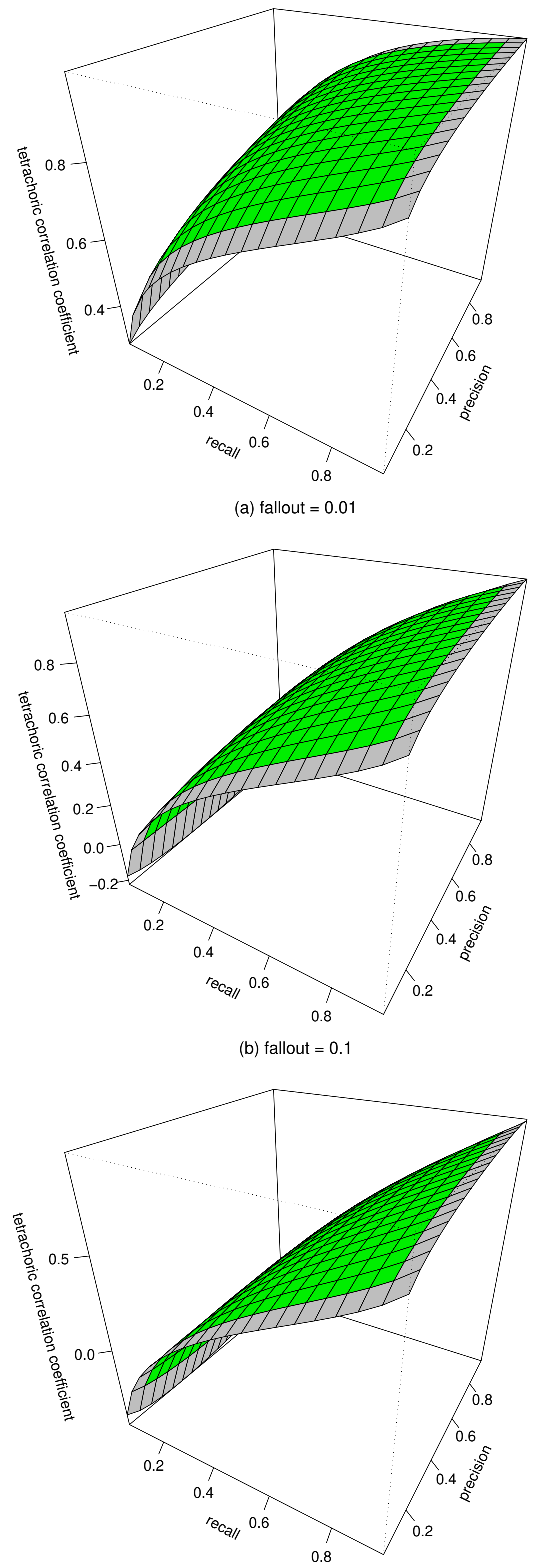
When $r = 0.4$ for example, the difference between ρ and ϕ becomes $(\rho - \phi) : 0.503 \to 0.381 \to 0.346$ according to $p : 0.05 \to 0.50 \to 0.95$. Since we know that both ρ and ϕ indicate a degree of association and can take a maximum value of 1, each difference is a quantity which equals half or one-third of the whole, and can be very a big value.

For comparison, Figure 3 shows the F measure of the z-axis to the same fallout a. Labels (a), (b), (c) discern between the weighting-factors of precision p to recall r, and correspond to $\beta = 0.5, 1.0, 2.0$. Since an F measure is a balance of r and p, (a) and (c) are in agreement with what replaced r and p, respectively.

We find that Figures 2 and 3 are alike, i.e., a phi coefficient and an F measure are similar (especially $\beta = 1$), when $a \leq 0.1$. For example, at $a = 0.01$ and $r = 0.4$, the difference varies to $(\phi - F_1) : -0.050 \to 0.009 \to 0.031$ according to $p : 0.05 \to 0.50 \to 0.95$. However, the degree of the approximation will worsen and it is impossible to say that they are the same above $a = 0.2$ as a increased in value. For example, at $a = 0.2$ and $r = 0.4$, the difference varies to $(\phi - F_2) : 0.159 \to 0.094 \to 0.276$ according to $p : 0.05 \to 0.5 \to 0.95$.

3.2 Comparison of break-even point with tetrachoric correlation coefficient and phi coefficient

Since the break-even point is the point at which the precision corresponds with the recall, for showing the tetrachoric correlation coefficient and the phi coefficient at this point, it is only necessary to show ρ and ϕ when cutting the curved surface with the plane of $r = p$ in Figures 1 and 2.

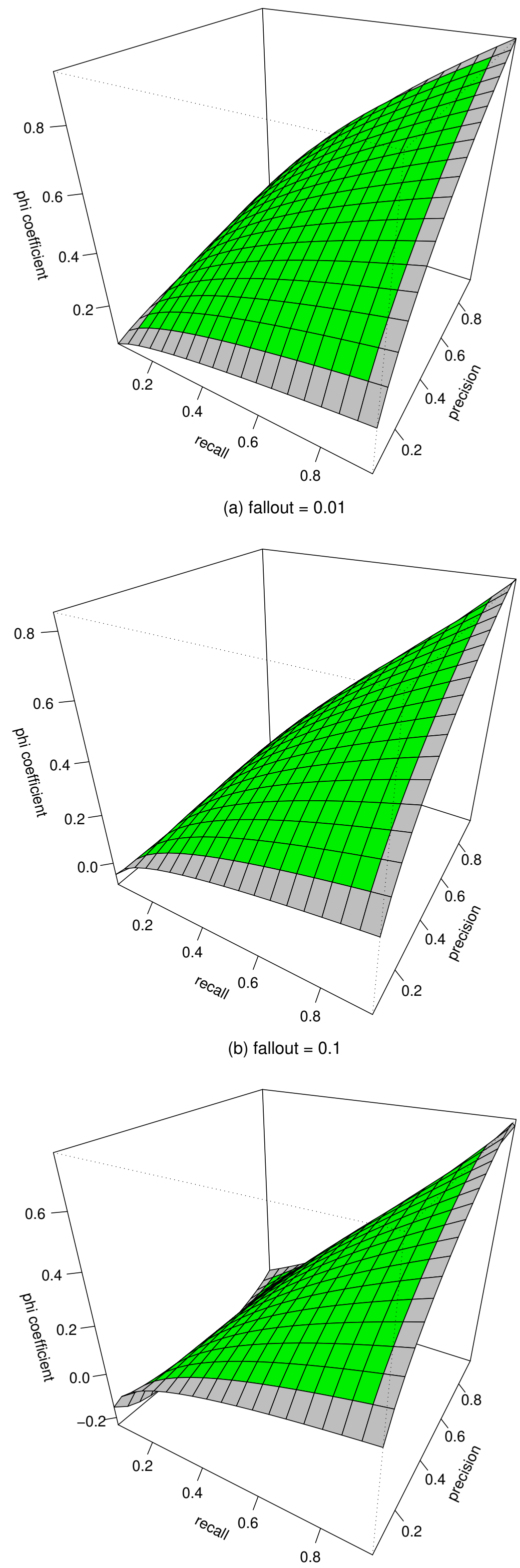


(a) fallout = 0.01

(b) fallout = 0.1

(c) fallout = 0.2

Figure 1. Tetrachoric correlation coefficient ρ

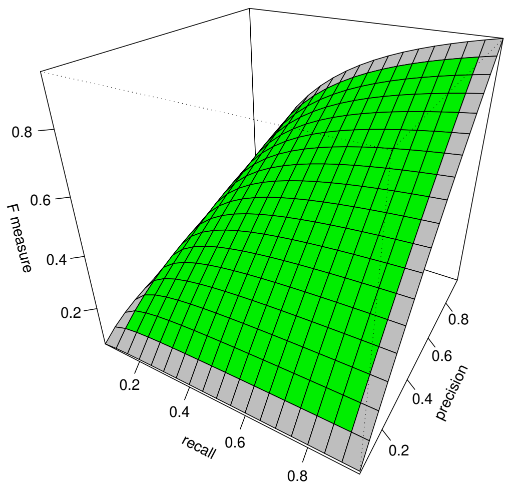


(a) fallout = 0.01

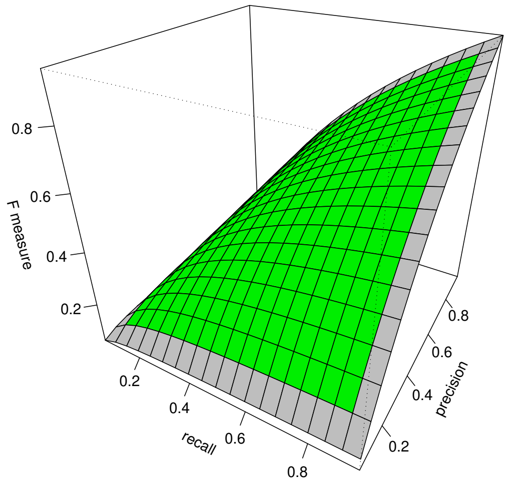
(b) fallout = 0.1

(c) fallout = 0.2

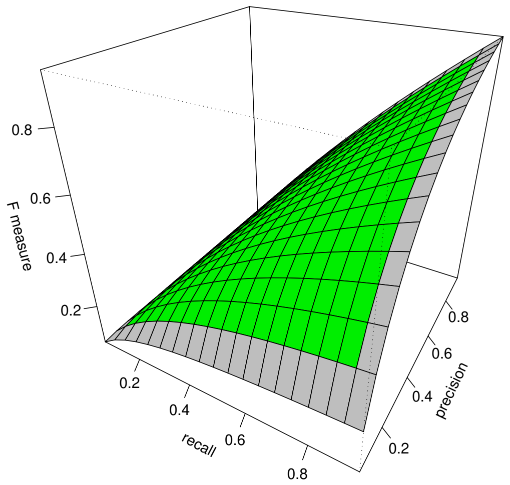
Figure 2. Phi coefficient



(a) beta = 0.5



(b) beta = 1



(c) beta = 2

Figure 3. van Rijsbergen's F-measure

Figures 4 and 5 show the tetrachoric correlation coefficient ρ and phi coefficient ϕ along the z-axis; recall or precision in the break-even point is shown on the x-axis, and fallout on the y-axis.

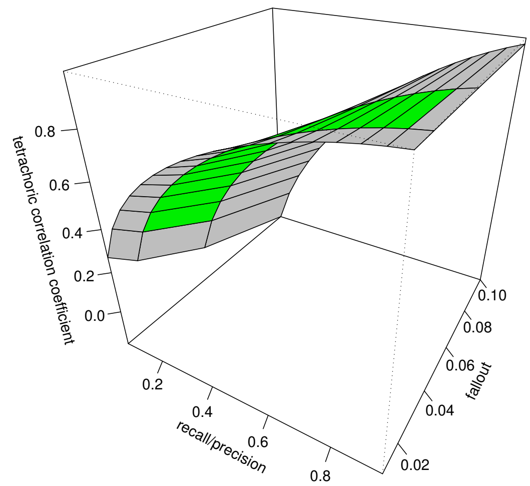


Figure 4. Break-even point and tetrachoric correlation coefficient

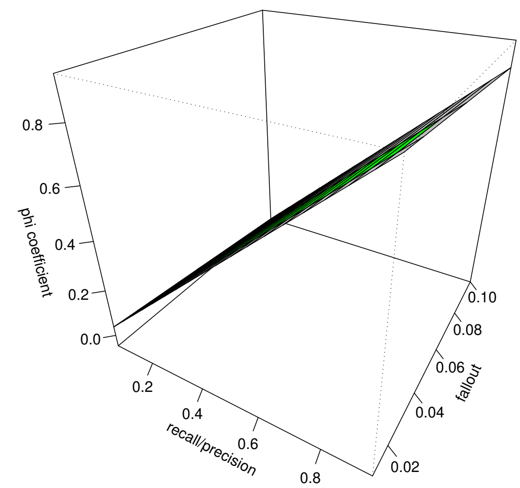


Figure 5. Break-even point and phi correlation coefficient

If the break-even point closely matches the tetrachoric correlation coefficient, the curved surface as drawn becomes a plane of $x = z$, which is the parallel displacement of the segment connecting $(x, z) = (0, 0)$ and $(x, z) = (1, 1)$ along the y axis. It is generally difficult to distinguish whether the curved surface is above or below from the plane of $x = z$. However, if the azimuth and complementary latitude (colatitude) of a viewpoint are equal, then the plane of $x = z$ from this viewpoint must be visible only to the segment which connects $(x, z) = (0.0)$ and $(x, z) = (1.1)$; it is very easy to distinguish a positional relation with the curved surface drawn. All the figures in this paper have an azimuths and a complementary latitudes of the same at 30 degrees.

Paying attention to this point and looking Figures 4 and 5 again, we find the curved surface in Fig. 4 is up from a plane of $x = z$ in all x-y coordinates, namely, it turns out that the tetrachoric correlation coefficient has acquired a larger value than that of the break-even point.

On the other hand, a phi coefficient is an index almost equivalent to the break-even point, so that we can say that the curved surface shown in Fig. 5 is just the state of the plane of $x = z$.

Indeed, a phi coefficient can be expressed using r, p, and a, as

$$\phi = (r - a)\sqrt{\frac{1-p}{r\left\{a + \left(\frac{1}{p} - 1\right)(1-a)\right\}}} \tag{14}$$

from a Eq. (1) – (3) and Eq. (6) – (9).

At the break-even point, since $r = p$, p can be eliminated. Since a is generally small in information retrieval, by setting $a \ll 1$, we can finally derive $\phi \approx r$.

3.3 Comparison of 11-point average precision with tetrachoric correlation coefficient and phi coefficient

Since 11-point average precision averages the precision at 11 standard recall levels of $(0.0, 0.1, \cdots, 1.0)$, in order to compare it, we need to calculate each p when changing into $r = 0.0, 0.1, \cdots, 1.0$ by giving a tetrachoric correlation coefficient ρ or phi coefficient ϕ and fallout a.

Although deriving p analytically is difficult, we can find p numerically by using a bisection method[6] etc. This method searches for p which let derived ρ and ϕ be the closest to their given values, considering the p's existential interval [0.01, 0.99].

However, as shown in Figures 1 and 2, ϕ is not necessarily uniform increased to p. Thus, p as obtained by the bisection method may be an approximation solution, but is not necessarily an optimum solution.

Regardless, we can obtain average precision $\bar{p}$ by taking the average at 11 points of p which can be determined numerically. Figures 6 and 7 scale it on z-axis, with ρ and ϕ on x-axis, and with a on y-axis.

Observing a vertical positional relation between the $x = z$ plane and the the curved surface drawn, we find from Fig. 6, a 11-point average precision $\bar{p}$ is a little smaller than the tetrachoric correlation coefficient ρ with the range of the fallout $a \leq 0.1$. For example, the difference varies to $\mathrm{Ave}(\bar{p} - \rho) : -0.21 \to -0.14 \to -0.014 \to -0.003$ according to $a : 0.01 \to 0.02 \to 0.05 \to 0.1$.

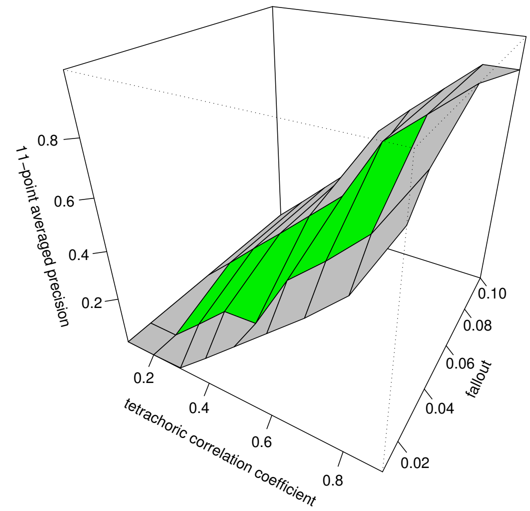


Figure 6. 11-point averaged precision and tetrachoric correlation coefficient

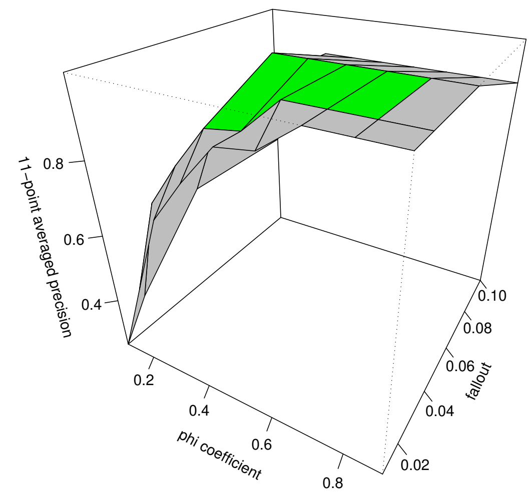


Figure 7. 11-point averaged precision and phi coefficient

We also find from Fig. 7, $\bar{p}$ takes a larger value than a phi coefficient and the difference varies to $\mathrm{Ave}(\bar{p} - \phi)$: $0.29 \to 0.35 \to 0.42 \to 0.34$ according to $a : 0.01 \to 0.02 \to 0.05 \to 0.1$.

It is important to note that the curved surface is not smooth. That is to say, z is not a convex for all x-y coordinates plane, because a bisection method were used to calculate of p.

4 Conclusion

This paper treats the range in which information retrieval is usually performed, namely, when fallout a is 0.1 or less. The evaluation criteria used by information retrieval systems and the related statistical indices are discussed. Furthermore, the quantitative mutual relationship between them has been clarified.

In the cross matrix in Table 1, when assuming a good retrieval, i.e., the retrieval with a sufficient recall and sufficient precision is performed, the frequency of f_{11} and f_{22} on the main diagonal line will become large, and the frequency of f_{12} and f_{21} on the remains will become small.

However, since typical information retrieval systems search a limited amount of data compared with the amount of data that are not searched, even if an element on the same main diagonal line is treated, the relation of $f_{11} \ll f_{22}$ is achieved in many cases.

An example of the $f_{11} \ll f_{22}$ relation is provided by the term "information retrieval"; the ratio is $f_{11}/f_{22} = 3.0 \times 10^{-3}$. Another example is given for the term "information retrieval & reliability"; the expression given is $f_{11}/f_{22} = 3.4 \times 10^{-5}$.

Therefore, we can easily expect that the criteria currently used to evaluate information retrieval systems are not necessarily similar to the related indices in the statistical 2×2 contingency table. I think that value is to have been able to grasp the the quantitative relationship between these indices.

Acknowledgement

The motivation behind this research is based on discussion with Prof. Carbonell at the Language Technologies Institute, Carnegie Mellon University between February, 2000 to December. During this stay, the author was a visiting researcher from Ministry of Education of Japan. The author would like to describe the above and to express his gratitude for their support.

References

[1] N.J. Belkin. The Problem of "matching" in Information Retrieval, Theory and Application of Information Research, 187–197, 1980.

[2] S. Fujita. More Reflections on "Aboutness" TREC-2001 Evaluation Experiments at Justsystem, NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, Maryland, November 13-16, 2001.

[3] J. Goldstein, M. Kantrowitz, M. Mittal, J. Carbonell. Summarizing Text Documents: Sentence Selection and Evaluation Metrics, 22nd International Conference on Research and Development in Information Retrieval, SIGIR 99, University of California, Berkeley, August 15–19, 1999.

[4] W. Ingwersen. *Information Retrieval Interaction*, Taylor Graham Publishing, London, 1993.

[5] D. E. Kraft, A. Bookstein. Evaluation of Information Retrieval System: A Decision Theory Approach, *Journal of the American Society for Information Science*, 29: 31–40, 1978.

[6] S.S. Rao. *Optimization, theory and applications*, A Holsted Press Book, John Wiley & Sons, 1979.

[7] G. Salton, C. Buckley. Term-weighting approaches in automatic text retrieval, Information Processing & Management, 24(5): 513–523, 1988.

[8] StatLib, Applied Statistics Algorithms, Available online. http://phase.etl.go.jp/stat/apstat/.

[9] K. Takeuchi.(ed.) *Dictionary in Statistics*, Toyo-keizai Press, 1989.

[10] TREC-8, Evaluation Techniques and Measures, TREC-8 Results, page A-1, NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8), Gaithersburg, Maryland, November 17-19, 1999. Available online. http://trec.nist.gov/pubs/trec8/t8_proceedings.html

[11] TREC-9, http://trec.nist.gov/pubs/trec9/appendices/A/web_results.html

[12] G.M. Quenot. TREC-10 Shot Boundary Detection Task: CLIPS System Description and Evaluation, 2001. NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, Maryland, November 13-16, 2001.

[13] C.J. van Rijsbergen. *Information Retrieval* (2 ed.) Butterworths, 1979.