# Information Retrieval System: Evaluation of Ranked Retrieval Results

Swapan Khan
Librarian
Narasinha Dutt College
Howrah
Email: khanswapan@gmail.com

Pronobi Porel
Librarian, Rabindra Mahavidyalaya, Hooghly
and Research Scholar, Deptt. of Library &
Information Science, Jadavpur University
Email: khanpronobi@gmail.com

*There are several measures used to evaluate an Information Retrieval System (IRS) to assess how well the search results satisfied the user's query intent. Ranked results are the core feature of an IR system. Precision, recall and F-measure are set-based measures, that cannot assess the ranking quality. Present paper shows to evaluate precision at every recall point which can also be used to evaluate a ranking of results of an IRS.*

***Keywords:*** *Information System, Recall, Precision, Information Retrieval, Ranked Retrieval, Evaluation of IRS*

## 1. Introduction

According to Wikipedia Information Retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. The evaluation of information retrieval systems is "the process of assessing how well a system meets the information needs of its users. There are two broad classe of evaluations: system evaluation and user-based evaluation. User-based evaluation measures the user's satisfaction with the system, while system evaluation focuses on how well the system can rank documents" (Voorhees, 2002). IR is interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, and statistics. IRS has developed as a highly empirical discipline. It is necessary to careful and through evaluation to exhibit performance of an IRS and to demonstrate its novel technique on document representations on response of a query.

## 2. Evaluation Series and Test Collections

There are so many evaluation series and test collections. The present study focuses on a list of the most standard test collections and evaluation series. These evaluation series mainly traced on test collections for ad hoc information retrieval system evaluation.

JOCAS
*Journal of Commerce, Arts and Science*
*Volume: 1, Issue: 1, June 2018*

### 2.1 The Cranfield Experiments:

Evaluation of IR systems is the result of early experimentation initiated by Cyril Cleverdon. He started a series of projects, called the Cranfield Projects, in 1957 that lasted for about 10 years in which he and his colleagues set the stage for information retrieval research (Heppin, 2012). This experiments was the pioneering test collection in allowing precise quantitative measures of information retrieval evaluation.

### 2.2 Text Retrieval Conference (TREC)

The National Institute of Standards and Technology (NIST), in 1992 has started a series of test bed for information retrieval. There were a lot number of test collections of different objects have been used in that project. The most important and popular one was TREC. The first TREC evaluation has been completed and evaluated between 1992 and 1999. The compositions of overall test collections were 6 CD ROMs containing 1.89 million documents and relevance judgments for 450 information needs which are called topics (Manning, Raghavan & Schutze, 2008). In 2003, to research in automatic segmentation, indexing, and content-based retrieval of digital video, Video Track has been performed which was called TRECVID. The TREC evaluation effort has grown in both the number of participating systems and the number of tasks each year. Ninety-three groups representing 22 countries participated in TREC 2003. In 2007, Genomics Tack was conducted to study the retrieval of genomic data, not just gene sequences but also supporting documentation such as research papers, lab reports, etc. Enterprise Track was done to study search over the data of an organization to complete some task in 2008. In recent, Clinical Decision Support Track, Contextual Suggestion Track, Dynamic Domain Track and etc. have been carried out (Text Retrieval Conference, 2017, October 4).

### 2.3 NII Test Collections for IR Systems (NTCIR)

It is a series of evaluation workshops designed to enhance research in information access technologies including information retrieval, question answering, text summarization, extraction, etc. to fulfill the following objectives (NTCIR Project Overview, n.d.):

- to encourage research in information access technologies by providing large-scale test collections.
- to present a forum on cross-system comparison and exchanging research ideas.
- to investigate evaluation methods of information access techniques and methods.

### 2.4 Cross Language Evaluation Forum (CLEF)

The Conference and Labs of the Evaluation Forum or CLEF, is an organization promoting research in multilingual information access "by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes" (Cross Language Evaluation Forum, n.d.). The CLEF organization arranged holds a conference every year in September in Europe since its first workshop in 2000.

## 3. Components of Evaluation

The main component to measure an IRS effectiveness is the combination of following three issues. These are

collectively called test collections:

- A document collection
- A set of expressible queries for information needs.
- A set of relevance judgments in binary mode of either relevant or non-relevant for each query-document pair.

Relevant and non-relevant documents retrieved from a test collection play a vital role in evaluation of IRS which refer to as the gold standard or ground truth of relevance. At the time of evaluation, it must be confirmed that the test collection and queries for information needs must be in reasonable in size. There must be fairly large test sets for average performance as results are highly variable over different documents and information needs. In general minimum 50 information needs has been accepted in this regard.

### 3.1 Recall and Precision:

To evaluate an IRS recall and precision are most widely used. Precision explains the exactness of the result of a search query and recall is used to show the completeness of the result of a search query. Both of them are widely used in statistical classifications. For evaluating recall and precision of an IRS, retrieved documents and relevance documents for a search query are considered. Recall is the measure of relevance documents retrieved over the total relevance documents where as precision is the ratio of relevance documents retrieved and total retrieved documents in a database. If IRS shows 100% relevance documents against a search query, it explains that a perfect precision score of 1.0 which means every result retrieved by a search was relevant. 100% recall defines that a perfect recall score of 1.0 which means all relevant documents were retrieved by the search. The results of a query in any IRS include one set of relevant documents and other set of non relevant documents. Following table shows their relationship:

Table 1: Analysis of search results by an IRS

| | Relevant | Non-relevant |
|---|---|---|
| Retrieved | true positive (tp) | false positive (fp) |
| Not retrieved | false negatives (fn) | true negative (tn) |

Sometimes relevant and non-relevant are defined by actual or true positive and actual or true negative respectively. On the other hand predictive positive and predictive negative denote the retrieved and not retrieved documents. Now recall and precision are explained as below:

$$\text{Re}call(R) = \frac{tp}{(tp + fn)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\text{Pr}ecision(P) = \frac{tp}{(tp + fp)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

### 3.2 F-Measure

To test the accuracy of attest F-measure is used which derived by Van Rijsbergen (1979). It considers both the precision (P) and the recall (R) of the test to compute the score (Evaluation in information retrieval, n.d.).. The traditional balanced F-score is the harmonic mean of recall and precision which is as follows (Khan & Bhattacharya, 2013):

$$F = 2 \cdot \frac{P \cdot R}{P + R} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The above formula is called F1 measure as recall and precision are evenly weighted.

The general formula of F-measure is which is described as below:

$$F_{\beta} = (1+\beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R} \quad where\ \beta > 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

The above formula can be expressed in terms of Type I and Type II errors as follows:

$$F_{\beta} = \frac{(1+\beta^2) \cdot tp}{[(1+\beta^2) \cdot tp + \beta^2 \cdot fn + fp]} \cdot Where\ \beta^2 \in [0, \infty] \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

### 3.3 Relation among Recall, Precision and F measure:

In any IRS it is fact that precision, recall and the F measure are set-based measures. Most of the search engines provide results which can newly be defined to extend these measures and to evaluate ranked retrieval results. In a ranked retrieval context, appropriate sets of retrieved documents are naturally given by the top retrieved documents.

### 3.4 Recall Precision Matrix (R-P Matrix):

Now from the following recall precision matrix F-measure provides significant results.

Table no. 2: Numerical presentation of output by an IRS by a single query

| | Relevant | Non-relevant | |
|---|---|---|---|
| Retrieved | tp = 30 | fp = 70 | tp + fp = 100 |
| Not retrieved | fn = 50 | tn = 40 | fn + tn = 90 |
| | tp + fn = 80 | fp + tn = 110 | tp + fp + fn + tn = 190 |

Values of R and P are ascertained from the above matrix where R = 0.375 and P = 0.30 [putting the value of tp, fp, fn in the above equation (1) and (2)].

### 4. Effects of F measure with R-P Matrix:

To measure $F_{\beta>1}$ i.e. in case of value of $\beta > 1$ the F-measure shows the result in below [in equation (5)]:

$$F_{\beta-2} = \frac{(1+4) \times 30}{[(1+4) \times 30 + 4 \times 50 + 70]} \quad where\ \beta = 2$$

$F_{\beta-2} = 0.357$ which emphasizes the recall (i.e. R = 0.375)

If $\beta < 1$, the value of [by putting $\beta = 0.5$ in equation (5)] will be as follows:

$$F_{\beta-0.25} = \frac{(1+0.25)\times 30}{[(1+0.25)\times 30 + 0.25\times 50 + 70]}$$

$F_{\beta-0.25} = 0.313$ which emphasizes the precision (i.e. P = 0.30)

**Explanation:**

$\beta$ is the determinant of recall or precision efficiency of an IRS where recall value 1 denotes a negligible precision results. If $\beta = 1$ then the formula (4) change into formula (3) which is the balanced F-score and called $F_{\beta-1}$. The value $1 < \beta < \infty$ in formula (5) emphasizes recall and value $1 > \beta > 0$ emphasizes precision (Khan & Bhattacharya, 2013).

**5. Ranked Retrieval**

Ranked results are the core feature of an IR system. Precision, recall and F-measure are set-based measures, that cannot assess the ranking quality. The solution is to evaluate precision at every recall point.

**5.1 Recall Precision Curve**

For each set of recall and precision values can be plotted to give a recall-precision curve as follows (Manning, Raghavan & Schutze, 2008).

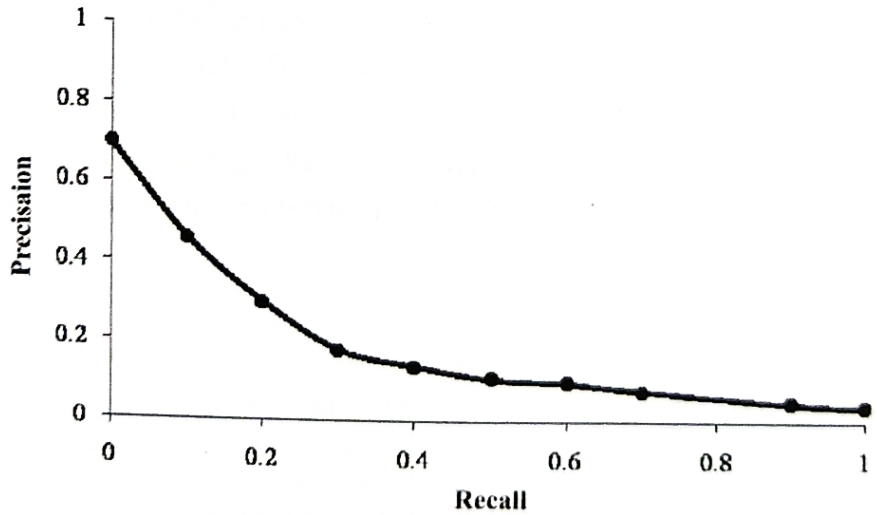


Fig.1: Averaged 11-point precision/recall graph across 50 queries for a representative TREC system.

In the graph, each recall level has been calculated the arithmetic mean of the interpolated precision at that recall level for each information need in the test collection. The TREC community, in recent, emphasis on Mean Average Precision (MAP), which provides a single-figure measure of quality across recall levels (Manning, Raghavan & Schutze, 2008). For a single information need, average precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs.

### 5.2 Mean Average Precision (MAP)

Average Precision is the Mean of the precision scores for a single query after each relevant document is retrieved. MAP means average precision measure, which measures the area underneath the entire recall-precision curve (Voorhees, 2002). Average of the precision value obtained for the set of top documents existing after each relevant document is retrieved, and this value is then averaged over information needs. That is, if the set of relevant documents for an information need $q_j \varepsilon Q$ is $\{d_1, \ldots\ldots\ldots\ldots d_{mj}\}$ and $R_{jk}$ is the set of ranked retrieval results from the top result until you get to document $d_k$, then (Manning, Raghavan & Schutze, 2008) the formula of MAP is as follows:

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}Precion(R_{jk})$$

When a relevant document is not retrieved at all the precision value in the above equation is taken to be 0. For a single information need, the average precision approximates the area under the uninterpolated precision-recall curve, and so the MAP is roughly the average area under the precision-recall curve (Zuva, & Zuva, 2012) for a set of queries.

## 6. Conclusion

Performance evaluation of IRS is vital at many stages in IRS development. At the final stage, this process it is important to show that how much a retrieval system achieves an acceptable level of performance. Therefore, in order to assess performance of a system it is essential to include some procedures which can be used to measure different stages of performance. Evaluation of ranked results of an IRS based on recall precision, suffers from practical disadvantages. In this study an indication has been shown towards measurement of performance of an IR system which shows either increasing or decreasing behaviour of recall or precision. The scalar measures of IRS are more popular as they give a definitive answer to which IRS is better and this measure gives an overall value of performance of the system.

## 7. References

Chowdhury, G.G. (1999). Introduction to modern information retrieval (2nd ed.). London: Facet Publishing.

Cross Language Evaluation Forum. ( n.d.). Retrieved January 22, 2018 from http://clef.isti.cnr.it/

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, Pittsburg, 2006. Retrieved September 11, 2017 from http://www.autonlab.org/icml_documents/camera-ready/030_The_Relationship_Bet.pdf

Deng, Y., Xu, J. and Gao, Y. (2008). Phrase table training for precision and recall: what makes a good phrase and a good phrase pair? Proceedings of Association for Computational Linguistics, Ohio, 81-88. Retrieved November 19, 2017 from http://www.aclweb.org/anthology/P/P08/P08-1010.pdf

Evaluation in information retrieval (Draft, Online edition). Cambridge University Press. Retrieved November 20, 2017, from http://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Evaluation in information retrieval (n.d.). Cambridge University Press. Retrieved July 20, 2017, from http://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Fleischhacker, D. and Stuckenschimidt, H. (2009). Implementing semantic precision and recall. Retrieved November 22, 2017 from http://www.dit.unitn.it/~p2p/OM-2009/om2009_poster9.pdf

Heppin, K. F. (2012). Test collections and the Carnfiedl Paradign. [Lecture 6]. Retrieved Janyary 20, 2018 from https://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/6IR12.pdf

Khan, S. & Bhattacharya, U. (2013). F-measure of an Information Retrieval System (IRS): Conceptual design for evaluation of unranked retrieval results. Librarian, 20(2), 7-10.

Lancaster, F. W. (1979). Information Retrieval systems: characteristics, testing, and evaluation (2nd ed.). New York: John Wiley.

Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R. (1999). Performance measures for information extraction. Proceedings of DARPA Broadcast News Workshop, Herndon. Retrieved December 25, 2017 from http://reference.kfupm.edu.sa/content/p/e/performance_measures_for_information_ext_114873.pdf

Manning, C. D., Raghavan, P. & Schutze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press. Retrieved anuary 15, 2018 from https://campus.fsu.edu/bbcswebdav/users/ bstvilia/lis5263IR/readings/08eval.pdf

NTCIR Project Overview (n.d.). Retrieved January 20, 2018 from http://research.nii.ac.jp/ntcir/outline/prop-en.html

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 311-318. Retrieved September November 22, 2017 from http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf

Prasher, R. G. (1989). Index and indexing systems. New Delhi: Medallion Press.

Text Retrieval Conference. (2017, October 4). Retrieved February 23, 2018 from https://en.wikipedia.org/wiki/Text_Retrieval_Conference

Turpin, A., Scholer, F. (2006). User performance versus precision measures for simple search tasks. Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval. Washington, 1118. Retrieved November 24, 2017 from http://researchbank.rmit.edu.au/eserv/rmit:2446/n2006001961.pdf

Van Rijsbergen, C. J. (1979). Information retrieval (2nd ed.). Newton, MA: Butterworth-Heinemann.

Voorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. Retrieved January 20, 2018 from https://www.inf.ed.ac.uk/teaching/courses/tts/handouts2017/VoorheesIREvaluation.pdf

Zou, K. and Hall, W. (2000). Two transformation models for estimating an ROC curve derived from continuous data. Journal of Applied Statistics, 27, 621-631.

Zuva, K. & Zuva, T. (2012). Evaluation of information retrieval systems. International Journal of Computer Science & Information Technology, 4(3), 35-43. doi: 10.5121/ijcsit.2012.4304

Journal of Commerce Arts and Science (JOCAS), June, 2018, Vol-1, No.-1, p. 53-59

# Information Retrieval System: Evaluation of Ranked Retrieval Results

Dr. Swapan Khan
Librarian, Narasinha Dutt College, Howrah
Email: khanswapan@gmail.com

Pronobi Porel
Librarian, Rabindra Mahavidyalaya, Hooghly and
Research Scholar, Deptt. of Library & Information Science, Jadavpur University
Email: khanpronobi@gmail.com

**Abstract:** There are several measures used to evaluate an Information Retrieval System (IRS) to assess how well the search results satisfied the user's query intent. Ranked results are the core feature of an IR system. Precision, recall and F-measure are set-based measures, that cannot assess the ranking quality. Present paper shows to evaluate precision at every recall point which can also be used to evaluate a ranking of results of an IRS.

**Keywords:** Information System, Recall, Precision, Information Retrieval, Ranked Retrieval, Evaluation of IRS

## 1 Introduction

According to Wikipedia Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. The evaluation of information retrieval systems is "the process of assessing how well a system meets the information needs of its users. There are two broad classes of evaluation, system evaluation and user-based evaluation. User-based evaluation measures the user's satisfaction with the system, while system evaluation focuses on how well the system can rank documents" (Voorhees, 2002). IR is interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, and statistics. IRS has developed as a highly empirical discipline. It is necessary to careful and through evaluation to exhibit performance of an IRS and to demonstrate its novel technique on document representations on response of a query.

## 2 Evaluation Series and Test Collections

There are so many evaluation series and test collections. The present study focuses a list of the most standard test collections and evaluation series. These evaluation series mainly traced on test collections for ad hoc information retrieval system evaluation.

### 2.1 The Cranfield Experiments:

Evaluation of IR systems is the result of early experimentation initiated by Cyril Cleverdon. He started a series of projects, called the Cranfield Projects, in 1957 that lasted for about 10 years in which he and his colleagues set the stage for information retrieval research (Heppin, 2012). This experiments was the pioneering test collection in allowing precise quantitative measures of information retrieval evaluation.

### 2.2 Text Retrieval Conference (TREC)

The National Institute of Standards and Technology (NIST), in 1992 has started a series of test bed for information retrieval. There were a lost number of test collections of different objects have been used in that project. The most important and popular one was TREC. The first TREC evaluation has been completed and evaluated between 1992 and 1999. The compositions of overall test collections were 6 CD ROMs containing 1.89 million documents and relevance judgments for 450 information needs which are called topics (Manning, Raghavan & Schutze, 2008). In 2003, to research in automatic segmentation, indexing, and content-based retrieval of digital video, Video Track has been performed which was called TRECVID. The TREC evaluation effort has grown in both the number of participating systems and the number of tasks each year. Ninety-three groups representing 22 countries participated in TREC 2003. In 2007, Genomics Tack was conducted to study the retrieval of genomic data, not just gene sequences but also supporting documentation such as research papers, lab reports, etc. Enterprise Track was done to study search over the data of an organization to complete some task in 2008. In recent, Clinical Decision Support Track, Contextual Suggestion Track, Dynamic Domain Track and etc. have been carried out (Text Retrieval Conference, 2017, October 4).

### 2.3 NII Test Collections for IR Systems ( NTCIR )

It is a series of evaluation workshops designed to enhance research in information access technologies including information retrieval, question answering, text summarization, extraction, etc. to fulfill the following objectives (NTCIR Project Overview, n.d.):

- to encourage research in information access technologies by providing large-scale test collections.
- to present a forum on cross-system comparison and exchanging research ideas.
- to investigate evaluation methods of information access techniques and methods.

### 2.4 Cross Language Evaluation Forum ( CLEF )

The Conference and Labs of the Evaluation Forum or CLEF, is an organization promoting research in multilingual information access “by (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes” (Cross Language Evaluation Forum, n.d.). The CLEF organization arranged holds a conference every year in September in Europe since its first workshop in 2000.

## 3 Components of Evaluation

The main component to measure an IRS effectiveness is the combination of following three issues. These are collectively called test collections:

- A document collection
- A set of expressible queries for information needs.
- A set of relevance judgments in binary mode of either relevant or non-relevant for each query-document pair.

Relevant and non-relevant documents retrieved from a test collection play a vital role in evaluation of IRS which refer to as the gold standard or ground truth of relevance. At the time of evaluation it must be confirmed that the test collection and queries for information needs must be in reasonable in size. There must be fairly large test sets for average performance as results are highly variable over different documents and information needs. In general minimum 50 information needs has been accepted in this regard.

**3.1 Recall and Precision:**

To evaluate an IRS recall and precision are most widely used. Precision explains the exactness of the result of a search query and recall is used to show the completeness of the result of a search query. Both of them are widely used in statistical classifications. For evaluating recall and precision of an IRS, retrieved documents and relevance documents for a search query are considered. Recall is the measure of relevance documents retrieved over the total relevance documents where as precision is the ratio of relevance documents retrieved and total retrieved documents in a database. If IRS shows 100% relevance documents against a search query, it explains that a perfect precision score of 1.0 which means every result retrieved by a search was relevant. 100% recall defines that a perfect recall score of 1.0 which means all relevant documents were retrieved by the search. The results of a query in any IRS include one set of relevant documents and other set of non relevant documents. Following table shows their relationship:

Table 1: Analysis of search results by an IRS

| | Relevant | Non-relevant |
|---|---|---|
| Retrieved | true positive (tp) | false positive (fp) |
| Not retrieved | false negatives (fn) | true negative (tn) |

Sometimes relevant and non-relevant are defined by actual or true positive and actual or true negative respectively. On the other hand predictive positive and predictive negative denote the retrieved and not retrieved documents. Now recall and precision are explained as below:

$$\mathrm{Re}\,call(R) = \frac{tp}{(tp + fn)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

$$\mathrm{Pr}\,ecision(P) = \frac{tp}{(tp + fp)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

**3.2 F-Measure**

To test the accuracy of attest F-measure is used which derived by Van Rijsbergen (1979). It considers both the precision (P) and the recall (R) of the test to compute the score (Evaluation in information retrieval, n.d.).. The traditional balanced F-score is the harmonic mean of recall and precision which is as follows (Khan & Bhattacharya, 2010):

$$F = 2 \cdot \frac{P \cdot R}{P + R} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The above formula is called $F_1$ measure as recall and precision are evenly weighted.

The general formula of F-measure is $F_\beta$ which is described as below:

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R} \quad where\ \beta > 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

The above formula can be expressed in terms of Type I and Type II errors as follows:

$$F_\beta = \frac{\left(1 + \beta^2\right) \cdot tp}{\left[\left(1 + \beta^2\right) \cdot tp + \beta^2 \cdot fn + fp\right]} \cdot \quad Where\ \beta^2 \in [0, \infty] \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

### 3.3 Relation among Recall, Precision and F measure:

In any IRS it is fact that precision, recall and the F measure are set-based measures. Most of the search engines provide results which can newly be defined to extend these measures and to evaluate ranked retrieval results. In a ranked retrieval context, appropriate sets of retrieved documents are naturally given by the top $k$ retrieved documents.

### 3.4 Recall Precision Matrix (R-P Matrix):

Now from the following recall precision matrix F-measure provides significant results.

Table no. 2: Numerical presentation of output by an IRS by a single query

| | Relevant | Non-relevant | |
|---|---|---|---|
| Retrieved | $tp = 30$ | $fp = 70$ | $tp + fp = 100$ |
| Not retrieved | $fn = 50$ | $tn = 40$ | $fn + tn = 90$ |
| | $tp + fn = 80$ | $fp + tn = 110$ | $tp + fp + fn + tn = 190$ |

Values of R and P are ascertained from the above matrix where $R = 0.375\ and\ P = 0.30$ [putting the value of tp, fp, fn in the above equation (1) and (2)].

### 5 Effects of F measure with R-P Matrix:

To measure $F_{\beta>1}$ i.e. in case of value of $\beta > 1$ the F-measure shows the result in below [in equation (5)]:

$$F_{\beta-2} = \frac{(1+4) \times 30}{[(1+4) \times 30 + 4 \times 50 + 70]}\ where\ \beta = 2$$

$F_{\beta-2} = 0.357$ which emphasizes the recall (i.e. $R = 0.375$)

If $\beta < 1$, the value of $F_{\beta<1}$ [ by putting $\beta = 0.5$ in equation (5)] will be as follows:

$$F_{\beta-0.5} = \frac{(1+0.25)\times 30}{[(1+0.25)\times 30 + 0.25\times 50 + 70]}$$

$F_{\beta-0.25} = 0.313$ which emphasizes the precision (i.e. $P = 0.30$)

**Explanation:**

$\beta$ is the determinant of recall or precision efficiency of an IRS where recall value 1 denotes a negligible precision results. If $\beta = 1$ then the formula (4) change into formula (3) which is the balanced F-score and called $F_{\beta-1}$. The value $1 < \beta < \infty$ in formula (5) emphasizes recall and value $1 > \beta > 0$ emphasizes precision (Khan & Bhattacharya, 2010).

## 6 Ranked Retrieval

Ranked results are the core feature of an IR system. Precision, recall and F-measure are set-based measures, that cannot assess the ranking quality. The solution is to evaluate precision at every recall point.

### 6.1 Recall Precision Curve

For each set of recall and precision values can be plotted to give a recall-precision curve as follows (Manning, Raghavan & Schutze, 2008).

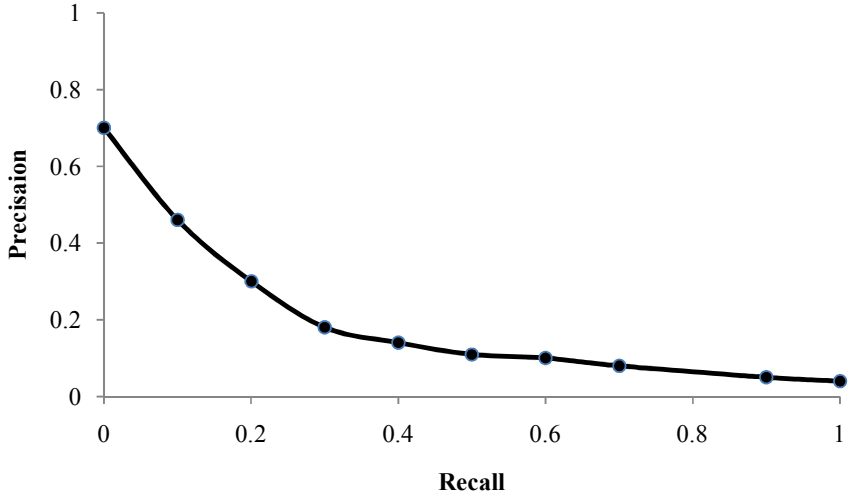


Fig.1: Averaged 11-point precision/recall graph across 50 queries for a representative TREC system.

In the graph, each recall level has been calculated the arithmetic mean of the interpolated precision at that recall level for each information need in the test collection. The TREC community, in recent, emphasis on Mean Average Precision (MAP), which provides a single-figure measure of quality across recall levels (Manning, Raghavan & Schutze, 2008). For a single information need, average precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs.

### 6.2 Mean Average Precision (MAP)

Average Precision is the Mean of the precision scores for a single query after each relevant document is retrieved. MAP means average precision measure, which measures the area underneath the entire recall-precision curve (Voorhees, 2002). Average of the precision value obtained for the set of top $k$ documents existing after each relevant document is retrieved, and this value is then averaged over information needs. That is, if the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, \dots \dots \dots \dots d_{mj}\}$ and $R_{jk}$ is the set of ranked retrieval results from the top result until you get to document $d_k$, then (Manning, Raghavan & Schutze, 2008) the formula of MAP is as follows:

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk})$$

When a relevant document is not retrieved at all the precision value in the above equation is taken to be 0. For a single information need, the average precision approximates the area under the uninterpolated precision-recall curve, and so the MAP is roughly the average area under the precision-recall curve (Zuva, & Zuva, 2012) for a set of queries.

## 7 Conclusion

Performance evaluation of IRS is vital at many stages in IRS development. At the final stage, this process it is important to show that how much a retrieval system achieves an acceptable level of performance. Therefore, in order to assess performance of a system it is essential to include some procedures which can be used to measure different stages of performance. Evaluation of ranked results of an IRS based on recall precision, suffers from practical disadvantages. In this study an indication has been shown towards measurement of performance of an IR system which shows either increasing or decreasing behaviour of recall or precision. The scalar measures of IRS are more popular as they give a definitive answer to which IRS is better and this measure gives an overall value of performance of the system.

## 8 References

Chowdhury, G.G. (1999). Introduction to modern information retrieval (2nd ed.). London: Facet Publishing.

Cross Language Evaluation Forum. ( n.d.). Retrieved January 22, 2018 from http://clef.isti.cnr.it/

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, Pittsburg, 2006. Retrieved September 11, 2017 from http://www.autonlab.org/icml_documents/camera-ready/030_The_Relationship_Bet.pdf

Deng, Y., Xu, J. and Gao, Y. (2008). Phrase table training for precision and recall: what makes a good phrase and a good phrase pair? Proceedings of Association for Computational Linguistics, Ohio, 81-88. Retrieved November 19, 2017 from http://www.aclweb.org/anthology/P/P08/P08-1010.pdf

Evaluation in information retrieval (Draft, Online edition). Cambridge University Press. Retrieved November 20, 2017, from http://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Evaluation in information retrieval (n.d.). Cambridge University Press. Retrieved July 20, 2017, from http://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Fleischhacker, D. and Stuckenschimidt, H. (2009). Implementing semantic precision and recall. Retrieved November 22, 2017 from http://www.dit.unitn.it/~p2p/OM-2009/om2009_poster9.pdf

Heppin, K. F. (2012). Test collections and the Carnfiedl Paradign. {Lecture 6]. Retrieved Janyary 20, 2018 from https://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/6IR12.pdf

Lancaster, F. W. (1979). Information Retrieval systems: characteristics, testing, and evaluation (2nd ed.). New York: John Wiley.

Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R. (1999). Performance measures for information extraction. Proceedings of DARPA Broadcast News Workshop, Herndon. Retrieved December 25, 2017 from http://reference.kfupm.edu.sa/content/p/e/performance_measures_for_information_ext_114873.pdf

Manning, C. D., Raghavan, P. & Schutze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press. Retrieved anuary 15, 2018 from https://campus.fsu.edu/bbcswebdav/users/bstvilia/lis5263IR/readings/08eval.pdf

NTCIR Project Overview (n.d.). Retrieved January 20, 2018 from http://research.nii.ac.jp/ntcir/outline/prop-en.html

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 311-318. Retrieved September November 22, 2017 from http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf

Prasher, R. G. (1989). Index and indexing systems. New Delhi: Medallion Press.

Text Retrieval Conference. (2017, October 4). Retrieved February 23, 2018 from https://en.wikipedia.org/wiki/Text_Retrieval_Conference

Turpin, A., Scholer, F. (2006). User performance versus precision measures for simple search tasks. Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval. Washington, 11–18. Retrieved November 24, 2017 from http://researchbank.rmit.edu.au/eserv/rmit:2446/n2006001961.pdf

Voorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. Retrieved January 20, 2018 from https://www.inf.ed.ac.uk/teaching/courses/tts/handouts2017/VoorheesIREvaluation.pdf

Zou, K. and Hall, W. (2000). Two transformation models for estimating an ROC curve derived from continuous data. Journal of Applied Statistics, 27, 621-631.

Zuva, K. & Zuva, T. (2012). Evaluation of information retrieval systems. International Journal of ComputerScience & Information Technology, 4(3), 35-43. doi: 10.5121/ijcsit.2012.4304