

Performance Evaluation of CLIR and MLIR using Precision Metric Variants

Pothula Sujatha
 Department of Computer science
 School of Engineering & Technology
 Pondicherry University
 Pondicherry, India
 spothula@gmail.com

Abstract

The performance evaluation of information retrieval systems has achieved a high momentum in the last few years. Basic performance measures of information retrieval systems include precision and recall. While these measures work well in monolingual web retrieval, they are not suitable for CLIR (Cross-lingual Information Retrieval) and MLIR (Multilingual Information Retrieval) where two or more languages are involved respectively. Many measures were proposed to improve over the precision-recall measures but they are inadequate to exhibit the language wise performance evaluations. Precision metric variants for evaluating the performance over the retrieval of the documents in various languages have been proposed in this research. This paper also identifies the major strengths and shortcomings of some of the existing IR performance evaluation measures. This paper concentrates on the metric based performance evaluation on two variants of IR. Experiments are conducted in two phases (CLIR and MLIR). These two phases of experiments have been done on practical web search systems and proved that the proposed measures are necessary to reveal the importance of language wise comparisons.

1. Introduction

Performance evaluation of any information retrieval (IR) system is very important task of the system development process. It is also mandatory part of the research process. Performance evaluation process has been a unifying feature of the IR research field. The performance evaluation of IR system tends to focus on either the user or the system. User-centered evaluation is crucial since it assesses the overall success of an IR system (Clough et al, 2008).

The performance evaluation of the IR system should also focus at the social level because more and more multilingual information is available on-line every day, i.e., documents are available in different languages and the user can retrieve the information in their native language easily. Traditional evaluation methodologies are inadequate to evaluate these

multilingual systems. So, new methodologies are required to evaluate them, in that motivation novel metrics are derived and proved that these novel metrics may be adequate to some extent to evaluate the performance of these systems.

After the traditional IR or monolingual IR, there are three kinds of IR variants are available in the IR research field. They are BLIR, CLIR and MLIR. In all these IR systems, the method and performance evaluation are same but the number of languages involved is different. In cross-lingual, two languages are involved, but in multilingual more than two languages are involved. In all these three IR systems query language is different from the document language/s.

CLIR means that the document set in a single language is searched for a topic in a different language, e.g., searching Japanese documents for German topics. Much of the research on CLIR has focused on the cases in which more than one translation is known for a query term. Ambiguity is an unavoidable consequence of using natural language, but CLIR applications must accommodate ambiguity in both the query language and the document language (Douglas, 2008).

There are many works are available in the literature on CLIR and MLIR. Many efficient systems of these kinds had been built to serve the need of finding information in different languages. The unique characteristic of MLIR system suggests specific strategies for evaluation. User cannot quickly read and evaluate many documents in a foreign language. Therefore, high precision should be an important goal for MLIR system. Once a few relevant documents have been collected, the system can resort to monolingual relevance feedback to find more relevant documents if high recall is the final goal.

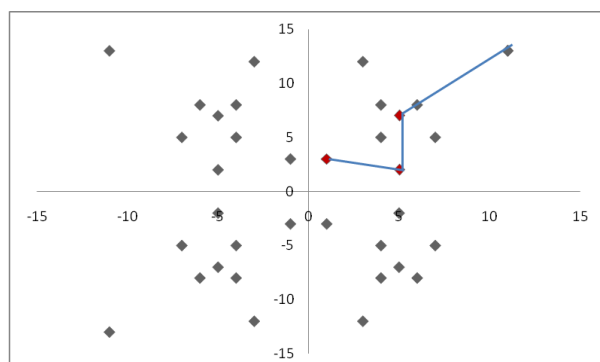


Figure 1. Database contexts Retrieval Nature of CLIR system

MLIR system can help the users to query in their native language and retrieve information in various foreign languages. Relevance was always the main concept for IR Evaluation. Relevance is a complex social and cognitive phenomenon. Therefore, Relevance among the languages involved in MLIR is also important to find. That is, how many relevant documents are retrieved in one language than the other languages involved in MLIR based on the source query language. For MLIR systems, analysis need to be done based on what factors this relevance difference is occurred. This analysis is based on two things: firstly, the collection is distributive or centralized, secondly, ranked retrieval or unranked retrieval. The results may vary once the source query language is changed regardless of the distributive or centralized document collection. The Figure 2 shows the performance aspects of the MLIR system. When any query in English is submitted to the retrieval system, the resultant documents are in French, English, Spanish and Telugu languages. The nature of the MLIR is to cover the documents in four languages. The line covers the relevant documents in the four document languages.

The language-specific modifications usually deal with issues such as word boundary determination, stop-list construction, stemming algorithms, etc. For some languages, however, these modifications can pose significant problems. For example, determining word boundaries is a difficult problem for Chinese. Similarly, devising stemming rules is likely to be a non-trivial problem for highly inflected languages like those used in South Asia. The treatment of compound words is also an issue for certain languages in which two or more words are combined into a *single* (non-hyphenated) word. On the whole, however, much of the research done by the IR community appears to be fairly language independent in nature.

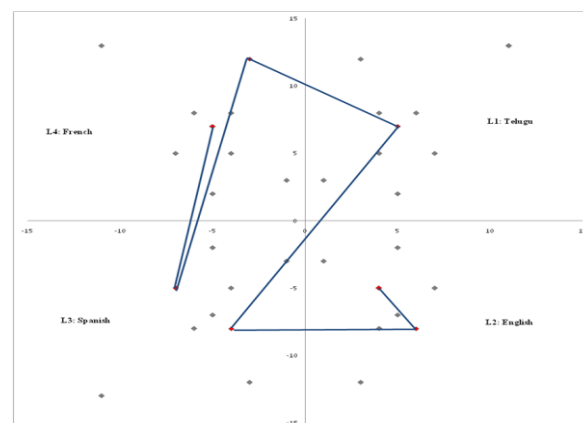


Figure 2. Retrieval nature of MLIR system

Additional problems in multi and cross language relevance assessment arise because of the added effort of performing topic creation and relevance assessment in multiple languages. A study is required for evaluation methodologies with respect to user needs in these systems. Very little is known as yet with respect to the expectations and real needs of the users of systems for MLIR. Even less is known as to how far the current evaluation infrastructure is really providing the best metrics to stimulate systems to meet these – as yet largely unknown – needs. This would be an important and valuable area of this research.

1.1. Drawbacks of existing Metrics

The systems are evaluated based on metrics of how well they retrieve the relevant documents and rank results. Examples of such metrics include Precision, Recall, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Precision at 10 (P@10), among many others (R. Baeza-Yates & G. Navarro, 1999). While these measures work well in closed-laboratory environments, they are not suitable for practical IR systems such as Web search systems. Evaluations based on precision and recall of topical queries may not only be difficult on the web, but incomplete (Craswell, 1999).

Many single-value measures were proposed to improve over the precision-recall measure, such as expected search length (ESL) (Cooper, 1968), average search length (ASL) (Losee, 1998) and Rank Power (Xiannong Meng, 2006). After this there are many other metrics have been identified by researchers e.g. (R.R. Korfhage, 1997). But these are based on the recall/precision measure, which presents the following problems. The first problem is: these measures are unable to present the ranks of retrieved documents explicitly. The next problem is, these measures are unable to work well in the web search systems and cannot practically identify and retrieve

all the documents that are relevant to a search query in the whole collection of documents. This is required by the recall/precision measure. The third problem is that it is not easy to read and interpret quickly what the measure means for ordinary users. The two other problems when using precision as measure of retrieval system performance are due to the weak ordering of output and the need for handling multiple queries (Vijay, 2003). In (Xiannong Meng, 2006), many single value measures such as F harmonic mean and E-measure have been identified to confront the third problem but these are not met the intended purpose of this paper. The drawback of f-score measure is that the number of documents to be retrieved is not fixed (Walid Magdy & Gareth, 2010).

The most vividly and generally used metric is MAP (Baeza-Yates, & Ribeiro-Neto, 1999) which emphasizes returning a greater number of relevant documents earlier. Since MAP is one of the precision variant measures which impacts on locating relevant documents later in the search of a ranked list is very frail, even if many such documents have been retrieved. For other types of IR task, the other IR evaluation metrics are found to be more representative than MAP. The metric MRR measures the performance when looking for one specific relevant item in a corpus (Azzopardi et al, 2007). MAP and GMAP (Geometric Mean Average Precision) are akin but GMAP is using geometric mean instead of the arithmetic mean. NDCG treats the relevant documents differently where the relevant documents are classified into classes according to the degree of relevance to the query. The objective is to find highly relevant documents earlier in the ranked list than less relevant ones.

In order to overcome shortcomings of MAP, the following measures like Bpref, inferred average precision (infAP), and rank-biased precision (RBP) are originated by the researchers. In (Moffat & Zobel, 2008) RBP is determined to produce a better modeling of user behavior in terms of how deep they are willing to go behind in the retrieved document list. Bpref and infAP are determined to conquer the problem of incomplete relevance judgments (Buckley & Voorhees, 2004). But infAP is used to collapses to MAP when judgments are complete (Aslam J. A., & E. Yilmaz, 2006). The three IR evaluation metrics are used to measure the effectiveness at retrieving relevant documents earlier rather than on the system recall. The f-score (Oard et al, 2008) combines recall with precision, and has been used for legal IR.

While this is sufficient and reasonable for monolingual IR focused systems, it is not suitable for systems where the objective is to find "all" relevant documents in one or more languages. Language wise effectiveness is also important to measure because documents are involved in many languages.

Evaluating these systems is essential with novel or modified metrics of monolingual IR.

For a multilingual document collection, good relevance feedback would probably necessitate obtaining at least one relevant document in each language of interest. So, precision is averaged at 5, 10, 15, and 20 documents retrieved as the evaluation measure for MLIR. The goal of the MLIR system is to obtain performance equivalent of its monolingual counterpart (Hull & Grefenstette, 1996). Language wise retrieval effectiveness is very important because the end user may know many languages. When the end user searching documents in MLIR system, if one language's retrieval effectiveness is more than the other language's then the end-user only looks at documents in this language in the same search engine since the other language document's retrieval effectiveness is lesser. That is, even though user knows many languages, he is interested in one particular language where he/she can gain more relevant documents than others. So it is very mandatory to know the language wise precision and recall values especially in MLIR systems. This is where MLIR systems are important and measuring its performance is important too.

2. Related work

In the literature, there are few papers regarding novel metrics for information retrieval. They are discussed in the following paragraphs:

A novel evaluation metric PRES have been devised for recall-oriented patent retrieval task (Walid Magdy et al, 2010). They also examine different evaluation measures for the same task and comparing different IR systems using scores. Families of metrics that only depend on the order of ranked items are rank-based metrics. The authors explored directly maximizing these metrics. These metrics allowed the authors to maximize different metrics for the same training data. (Donald et al, 2005) There are metrics which are optimized based on some smooth approximation with gradient descent; they are NDCG and AP. In (Olivier et al, 2010), an annealing algorithm has been proposed which was designed to optimize these two measures. Their main idea is to minimize a smooth approximation of these two measures with gradient descent. They have provided theoretical analysis on the choice of smoothing factor.

An overview of the current activities of the major evaluation initiatives have been given in (Thomas Mandl, 2008). Special attention is given to the current tracks and developments within TREC, CLEF and NTCIR. There are two different architectures in MLIR: centralized and distributed. In (Evangelin et al, 2008), they authors simulated the performance

metrics of an IR system in these two different architectures. They also presented a procedure to calculate the response time early in the life cycle. Many single-value measures were proposed to improve over the precision-recall measure, such as expected search length (ESL), average search length (ASL) and Rank Power. Authors compared in this paper the measures of ESL, ASL, and Rank Power applied to a set of real Web retrieval data. In (Xiannong Meng, 2006), the results demonstrate that Rank Power indeed is a feasible, effective, and easy to use single-value measure for performance of practical IR systems such as Web search engines.

The recent research has suggested an evaluating IR systems based on user behavior. The effectiveness of IR is usually evaluated using Normalized Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP) and Precision at K on a set of judged queries. In this paper, they have elaborated about the experiments that interleave two rankings and track user clicks. A study on interleaving was discussed in (Filip Radlinski & Nick Craswell, 2010), when comparing it with traditional measures in terms of reliability, sensitivity and agreement. Here, the authors stated that the interleaving experiments can identify large differences in retrieval effectiveness with much better reliability than other click-based methods. They have concluded that amongst the traditional measures NDCG has the strongest correlation with interleaving. At last, they also described an approach to enhance interleaving sensitivity with some new forms of analysis. A comparison between MAP and GMAP through t-test is given in (G.V. Cormack & T.R. Lynam, 2007). They have examined not only t-test, but also wilcoxon test and sign test in finding the difference between two IR systems is important or not. All these tests performed on subsets of the TREC 2004 Robust Retrieval collection.

3. Metrics for CLIR System evaluation

Metrics are chosen by the retrieval system designer based on the underlying task. The ultimate goal of the retrieval system is then to maximize the chosen metric/s. This is often accomplished by hand tuning system parameters until a given performance level is achieved (Donald et al, 2005). From this above view point, the importance of the metrics when one or more languages are involved in the retrieval system is known. In this paper, new precision metrics for CLIR are devised and measured through experiments taken from the web. This paper describes a study analyzing the behavior of available evaluation metrics when applied to variants of IR systems. The results of this analysis are used to motivate the proposal of a novel evaluation metrics

which modifies the existing IR metrics particularly, Precision.

Precision Metrics

Precision, measures (Olson et al, 2008) how many of the documents retrieved are actually relevant. The standard measure for monolingual precision (η) is defined as follows:

$$\eta = \frac{tp}{tp+fp} \quad (1)$$

where, tp represents true positives i.e. relevant documents from the retrieved document results

fp represents false positives i.e. relevant documents that are missed from the retrieved results

It is a general measure, which gives the relevant documents among the retrieved documents. It may be varied from search engine to search engine. Basically, it has been derived for monolingual IR purposes later it also used for CLIR Systems. This metric might also varied from language to language when two languages are involved in the retrieval system i.e. CLIR. Variance will be there because the total number of relevant documents available in each language is differ. The number of documents retrieved in each language is also not identical in these systems. Hence, performance is different from one system to another though same languages are involved. So, it is obvious to know the language wise precision values over the languages. Therefore, the Eq. (1) can be modified for CLIR systems as follows:

$$\eta_{(1)} = \frac{tp_{(2)}}{tp_{(2)}+fp_{(2)}} \quad (2)$$

where, tp represents true positives i.e. relevant documents in language 2 from the retrieved document results

fp represents false positives i.e. relevant documents in language 2 that are missed from the retrieved results

This work, used the standard precision metric, and manually verifies how many documents are retrieved and relevant in each language. But it is expensive and time taking process. Many of the researchers used the traditional precision metric in evaluating performance of the MLIR systems.

The Eq. (2) is important with the stated reason in the above paragraph. Here 1 specifies the language involved in the evaluation process.

$$\eta_{(2)} = \frac{tp_{(1)}}{tp_{(1)}+fp_{(1)}} \quad (3)$$

where, tp represents true positives i.e. relevant documents in language 1 from the retrieved document results

fp represents false positives i.e. relevant documents in language 1 that are missed from the retrieved results

The above Eq. (3) is used to give the precision value for the language 2. If two languages i.e. $n = 2$ are involved then CLIR systems performance is evaluated with $\eta(1)$ and $\eta(2)$.

$$\eta R_{(1)} = \frac{\eta(1)}{\eta(2)} \quad (4)$$

Eq. (4) is used to see the comparative performance level of precision of language-1 over language-2. Eq. (5) would reveal the comparative performance level of precision of language-2 over language-1.

$$\eta R_{(2)} = \frac{\eta(2)}{\eta(1)}$$

$$\eta U_{(1)} = \frac{(\eta(1) - \eta(2))}{\eta(2)} \quad (6)$$

$$\eta U_{(2)} = \frac{(\eta(2) - \eta(1))}{\eta(1)} \quad (7)$$

If $\eta U_{(1)}$ or $\eta U_{(2)}$ gives positive value then it is literally upward performance/better performance otherwise it is literally downward performance/bitter performance. When the comparative level of precision between two languages reaches the average or not is checked using the Eq. (6) and (7). If $\eta U_{(1)}$ gives negative value then the retrieved documents in language-1 are very less important, that is, non-relevant documents are more to examine. If $\eta U_{(2)}$ gives negative value then the retrieved documents in language-2 are very less important, that is, non-relevant documents are more to examine.

4. Metrics for MLIR System evaluation

Let $L = \{L_1, L_2, L_3, \dots, L_N\}$, where L = retrieved document language and N = No. of languages involved in retrieval system.

$$\eta(i) = \frac{tp(i)}{tp(i) + fp(1)} \quad (8)$$

where $i = 1$ to N ,

tp represents true positives i.e. relevant documents in language i from the retrieved document results

fp represents false positives i.e. relevant documents in language i that are missed from the retrieved results

If ' N ' languages are involved in the retrieval system then $\eta(1), \eta(2), \dots, \eta(i)$ measures have to be evaluated for MLIR system and ' i ' can be extended to the desired level.

The same metrics given in Eq. (2) to (7) of CLIR are applied to MLIR with minor changes in the place of 1 put ' i ' and in the place of 2 put ' j ' especially when comparing precision values among languages involved in retrieval process.

Yet another metric called normalized performance/precision is derived for the purpose of normalized value of the precision over ' N ' languages which are given in Eq. (9).

Normalized performance of precision:

The normalized performance of the precision values related to multiple languages is given below:

$$\eta N_{(i)} = \frac{\eta(i)}{\max(\eta\{L\})} \quad (9)$$

For example, if the retrieval system involved in 8 languages then $L = \{L_1, L_2, L_3, \dots, L_8\}$. The precision-normalized for L_3 can be defined as in Eq. (10) as follows:

$$\eta N_{(3)} = \frac{\eta(3)}{\max(\eta\{L\})} \quad (10)$$

That is, $\max(\eta\{L\}) = \max(L_1, L_2, L_4, L_5, L_6, L_7, L_8)$ and excluding L_3 .

5. Experimentation and Result Analysis

There are variants presented in the literature for standard precision. Precision variants are: MAP, P@5-15, R-precision, AP, GMAP etc. None of these exhibit performance in terms of individual language that is involved. Derivation of novel precision metric variants supposed to do a language wise performance evaluation. That is, $\eta(1)$ with respect to language-1 and $\eta(2)$ with respect to language-2. This kind of evaluation is essential where we need to concentrate on the importance of standard and native languages. For example, a search engine which can be used to get documents in Telugu and English, then much importance is given to standard language English than Native language Telugu. This is happened in almost all search engines. Since, performance of the systems is ignored due to less importance.

MLIR systems are having many characteristics/attributes. These characteristics/attributes are mapped with available metrics. This paper concentrates on mapping the characteristics/attributes with standard precision metrics and the proposed precision metric variants. The following properties are mapped with precision are: Retrieval effectiveness, robustness, relevancy, efficiency and consistency.

The following properties are mapped with proposed precision metric variants apart from the attributes mapped with standard precision: language flexibility, language model probability, translation

model probability, translation ambiguity, degree of ambiguity, verbosity of the language involved, language relatedness and indexing facility. The use and the results of precision metric variants applied against a set of Web search results. The experimental data was collected by sending chosen queries to the below specified search engines.

Experimental Phases

The experiments are described in two Phases. Phase I is used to assess the performance of CLIR system and the Phase II is used to assess the performance evaluation of MLIR system.

Performance Measures

The proposed Precision metrics are used to assess the Phase I and Phase II Experiments. Performance is measured with phase I experiments with phase II experiments.

Test environment

The phase I and II are tested and evaluated using five search engines: they are Google, Yahoo, Altavista, Bing and Ask. The languages involved in these experiments are French (F), German (G), Spanish(S), Italian (I) and Dutch (D). Each language query is translated using Babylone fish and Systran translators. Based on these query translations, the translated query is submitted to the above specified search engines and the retrieved documents in various languages are evaluated.

The measured values of proposed precision variants are calculated for Google, Alta vista, Bing, Ask and Yahoo search engines and shown in Table 1, Table 2, Table 3, Table 4 and Table 5 respectively. Phase I results gives language wise comparisons which are applied to the web search data. These observations exhibit the effectiveness of the enhanced basic measures. Table 6 demonstrates the MLIR retrieval effectiveness among the five search engines. The normalized precision performance is measured and given. For example F is the query language and the corresponding document languages are G, S, I and D.

Table 1. Performance results of Precision Metric variants (Google)

CLIR				
Run	$\eta R_{(1)}$	$\eta R_{(2)}$	$\eta U_{(1)}$	$\eta U_{(2)}$
F-G	0.0625	0.032	0.0125	-0.008
S-I	0.0404	0.005	0.016	-0.005
G-S	0.079	0.0102	0.039	-0.00495
I-F	0.002	0.247	-0.008	0.019
D-G	0.0625	0.032	0.0125	-0.01

Table 2. Performance results of Precision Metric variants (Alta Vista)

CLIR				
Run	$\eta R_{(1)}$	$\eta R_{(2)}$	$\eta U_{(1)}$	$\eta U_{(2)}$
F-G	0.039	0.014	0.011	-0.285
S-I	0.02	0.02	0	0
G-S	0.02	0.02	0	0
I-F	0.014	0.039	-0.005	0.4
D-G	0.02	0.02	0	0

Table 3. Performance results of Precision Metric variants (Bing)

CLIR				
Run	$\eta R_{(1)}$	$\eta R_{(2)}$	$\eta U_{(1)}$	$\eta U_{(2)}$
F-G	0.048	0.012	0.0170	-0.007
S-I	0.04	0.005	0.02	-0.005
G-S	0.02	0.02	0	0
I-F	0.003	0.096	-0.0067	0.651
D-G	0.02	0.02	0	0

Table 4. Performance results of Precision Metric variants (Ask)

CLIR				
Run	$\eta R_{(1)}$	$\eta R_{(2)}$	$\eta U_{(1)}$	$\eta U_{(2)}$
F-G	0.057	0.022	0.0150	-0.008
S-I	0.036	0.055	-0.005	0.007
G-S	0.022	0.057	-0.008	0.015
I-F	0.055	0.036	0.0069	-0.005
D-G	0.012	0.048	-0.007	0.017

Table 5. Performance results of Precision Metric variants (Yahoo)

CLIR				
Run	$\eta R_{(1)}$	$\eta R_{(2)}$	$\eta U_{(1)}$	$\eta U_{(2)}$
F-G	0.093	0.427	0.03	-0.013
S-I	0.25	0.002	0.04	-0.008
G-S	0.32	0.062	-0.008	0.0125
I-F	0.001	0.374	-0.008	0.313
D-G	0.022	0.053	-0.007	0.0133

Table 5. Normalized precision performance

MLIR normalized precision $\eta N_{(i)}$					
Lang.	GOOGLE	YAHOO	ALTA VISTA	ASK	BING
F	0.7911	0.2906	0.0576	0.0590	0.0760
S	0.5113	0.7812	0.8333	0.6315	0.5128
G	0.099	0.4096	0.4166	0.3859	0.5128
I	0.0253	0.0031	0.0625	0.9649	0.3589
D	0.7911	0.0687	0.4166	0.2105	0.5128

5.1. Result analysis

We can draw the following observations from the data. Here our focus is not on the search engines but on the enhanced precision measure's language wise comparisons. In phase I (CLIR performance): Relative precision in language-1 ($\eta R_{(1)}$) is more between the G-S run and it is very low with I-F and D-G language runs. In the same way, the Relative precision in language-2 ($\eta R_{(2)}$) is more between the F-G run and it is very low with S-I language run. In all most all the search engines the S-I run is having poor performance. The possible reasons are: the documents in I relevant to S query or very less. There may be ambiguity in query formulation and translation. May be less number of documents in I and. Among them very few are relevant.

The run I-F is always having the bitter performance when compared to other runs except in Table 4. The other CLIR runs are having better performance (since no negative value with $\eta U_{(1)}$). On the other side, the $\eta U_{(2)}$, is giving bitter performance with the run F-G all the web results. Much Better performance (positive and larger value) is demonstrated by the document language F amongst the five runs of precision up measure except the run in Ask (bitter performance).

In Phase II (MLIR performance), the same result analysis can be done between multiple language. Apart from these, the normalized precision performance is measured with five languages. The language I is better in performance rather than the other languages which are involved in MLIR retrieval. After that language S is better. As we described above, the retrieval performance between systems to systems is varied. The language G is having the average performance. That is, in all the retrieval systems G documents are marginal. From this, we conclude that retrieval system be effective and efficient in terms of good document collection, good indexing system and finally good ranking model.

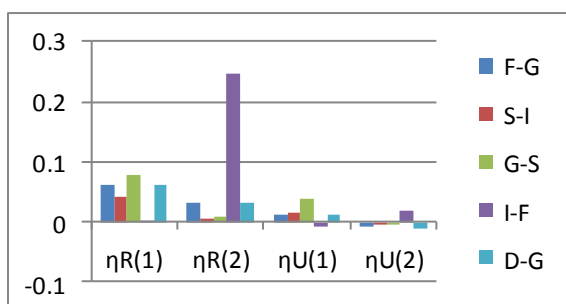


Figure 3 Performance evaluations of Precision metric variants in Google

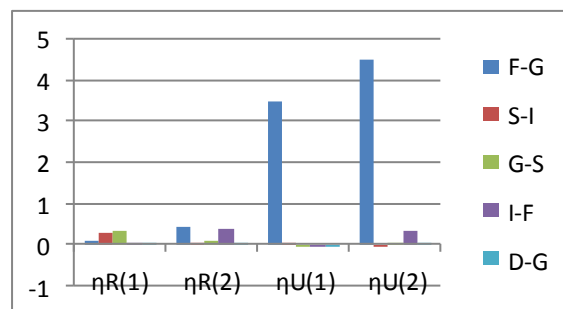


Figure 4 Performance evaluations of Precision metric variants in Yahoo

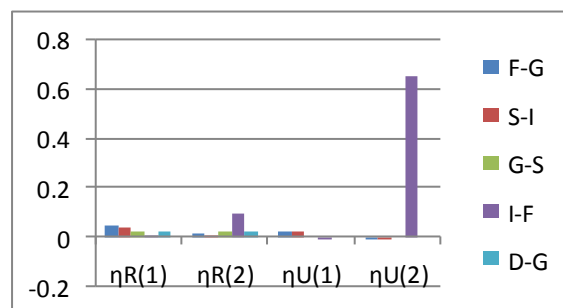


Figure 5 Performance evaluations of Precision metric variants in Alta vista

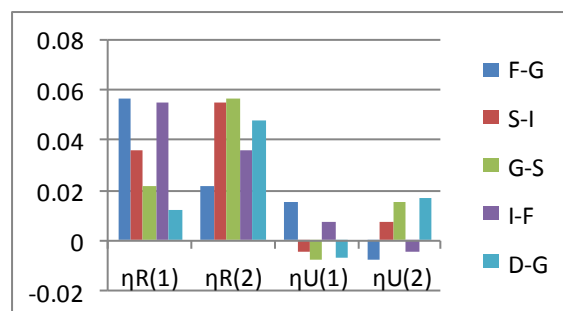


Figure 6 Performance evaluations of Precision metric variants in Ask

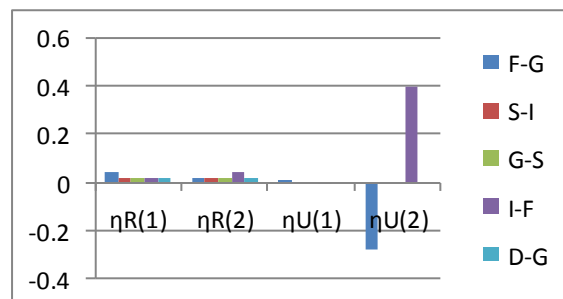


Figure 7 Performance evaluations of Precision metric variants in Bing

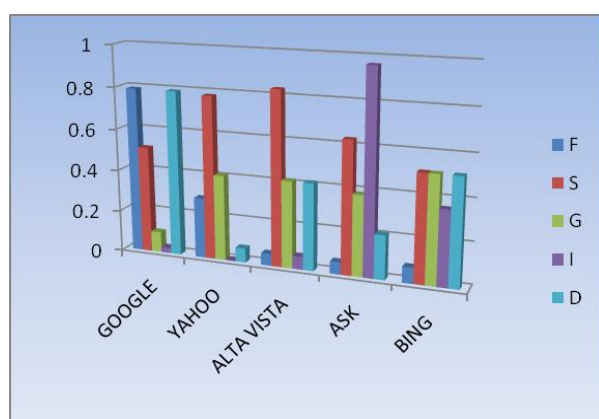


Figure 8 MLIR normalized performance

6. Conclusion

This paper presented the enhanced versions of the basic measure i.e. Precision metric variants. It also enumerated the advantages and disadvantages of the traditional measures. The new measures are derived for CLIR and MLIR systems. Because in these systems the language wise comparisons are not easy with traditional IR Precision measures. New measures are suitable to evaluate the performance of the retrieval systems where more than two languages are involved. The queries translated and submitted to the web search engines. The experimental results were taken for CLIR and MLIR. Performance evaluation is done among the web search engines using the proposed measures. The comparisons done in result analysis may not be possible with standard Precision metric.

High precision is tedious to achieve. This effect is applicable when two or more languages are involved in the same CLIR and MLIR system. Researchers are trying to achieve the same performance level of monolingual IR with CLIR/MLIR. But that has many practical problems involved. Because of language relatedness, importance of the language for the user when much language is involved, tools of the languages (query or document translation) may be different, etc. So, the precision analysis over the languages is very significant in the new era.

In future, concentration diverted towards on recall metric variants and other measures like F-measure, P@R and MAP etc. we will perform the experiments with one of the existing evaluation initiatives data.

7. References

- [1] A.B. Smith, C.D. Jones, and E.F. Roberts, "Article Title", *Journal*, Publisher, Location, Date, pp. 1-10.
- [2] Jones, C.D., A.B. Smith, and E.F. Roberts, *Book Title*, Publisher, Location, Date.
- [1] Aslam J. A., and E. Yilmaz. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of CIKM '06*, page 102-111, 2006.

- [2] Azzopardi L., de Rijke, M., and K. Balog. Building simulated queries for known-item topics: an analysis using six European languages. In *Proceedings of SIGIR '07*, pages 455-462, 2007.
- [3] Baeza-Yates, J., and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley, 1999.
- [4] R. Baeza-Yates and G. Navarro. *Modern Information Retrieval*. Addison-Wesley, 1999.
- [5] Buckley, C., and Voorhees, E. M. Retrieval evaluation with incomplete information. In *Proceedings of SIGIR '04*, pages 25-32, 2004.
- [6] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of ACM SIGIR*, pp. 33-40.
- [7] P. Clough, J. Gonzalo, J. Karlgren, E. Barker, J. Ariles, V. Peinado (2008), "Large-Scale Interactive Evaluation of Multilingual Information Access Systems - the iCLEF Flickr Challenge", 30th European Conference on Information Retrieval (ECIR 2008), pp. 33-38.
- [8] W.S. Cooper, Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems, *Journal of the American Society for Information Science*, 19(1), 30-41, 1968.
- [9] G.V. Cormack and T.R. Lynam, "Validity and power of t-test for comparing MAP and GMAP", in *Proc. SIGIR*, 2007, pp.753-754.
- [10] Craswell, N., P. Bailey, and D. Hawking. "Is it fair to evaluate web systems using trec ad hoc methods?." In *ACM Conference on Research and Development in Information Retrieval*, Berkeley, CA, 1999
- [11] Donald A. Metzler, W. Bruce Croft, and Andrew McCallum: Direct Maximization of Rank-Based Metrics for Information Retrieval. July 16, Published at CIIR 2005.
- [12] Douglas W. Oard, Daqing He, and Jianqiang Wang: User-assisted query translation for interactive cross-language information retrieval. *Inf. Process. Manage.* 44(1): 181-211 (2008)
- [13] D Evangelin Geetha G Krishna Naidu T V Suresh Kumar K Rajani Kanth, "Simulative Performance Evaluation of Information Retrieval Systems" *Second Asia International Conference on Modelling & Simulation*, 2008, pp 297-302.
- [14] Filip Radlinski, Nick Craswell: Comparing the sensitivity of information retrieval metrics, *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, July 2010, 667-674.
- [15] Hull, D. A., & Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*.
- [16] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of ACM SIGIR*, pp. 41-48.
- [17] R.R. Korfhage, *Information Storage and Retrieval*, John Wiley & Sons, 1997.
- [18] R.M. Losee, *Text Retrieval and Filtering: Analytic Models of Performance*, Kluwer Publisher, Boston, 1998.
- [19] Mandar Mitra, B. B. Chaudhuri: *Information Retrieval from Documents: A Survey*. Volume 2, Numbers 2/3, May 2000, 141-163