PERFORMANCE EVALUATION OF CROSS AND MULTILINGUAL INFORMATION RETRIEVAL USING RECALL METRIC VARIANTS

Pothula Sujatha
Department of Computer science, School of Engineering & Technology, Pondicherry University, Pondicherry, India

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 and recall measures but they are inadequate to exhibit the language wise performance evaluations. Recall 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.

KEYWORDS
Evaluation; Cross-Lingual Information Retrieval; Performance Evaluation; Characteristic

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 recall 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.

1.1. Types of Information Retrieval (IR) 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 recall 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.

1.2. Systems Importance of performance evaluation of CLIR and MLIR Systems

The CLIR methods are useful for people who can read a certain language, but they do not know enough how to construct effective queries in that language (Mandar Mitra & Chaudhuri, 2000). The nature of CLIR system is depicted in Fig. 1. The query language is English and the resultant document set is in Telugu language. Since it is a CLIR system, the documents covered in only one language are showed. The line covers the relevant documents in the document language (result set).

Figure 1. 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, the number of relevant documents are retrieved in one language is different from 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 Fig. 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.

1.3. 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. Korrhage, 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 (Axlam 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 either 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. 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 recall values especially in MLIR systems. This is where MLIR systems are important and measuring its performance is important too.

2. Related work

The following formatting rules must be followed strictly. This (.doc) document may be used as a template for papers prepared using Microsoft Word. Papers not conforming to these requirements may not be published in the conference proceedings.

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. In this paper authors compared 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. 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, recall metric variants 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, recall.

Recall Metric variants
The proposed metrics reveal the most important credential of the retrieval systems where more than two languages are involved.

CLIR measures
Recall measures how many of the relevant documents in a collection have actually been retrieved. The standard measure for monolingual recall (\( \Theta \)) is defined (Olson et al, 2008) as follows:

\[
\Theta = \frac{tp}{tp+fn}
\]

(1)

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

\( fn \) 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 actually relevant documents. As precision values, Recall values also may be varied from search engine to search engine. Basically, it has been used for monolingual IR purposes. The recall metric might also varied from language to language when many languages are involved in the retrieval system i.e. MLIR. Because the total number of relevant documents available in each language is differ. Hence, performance is different from one MLIR system to another MLIR system though same languages are involved. So, it obvious to measure the recall values over the languages involved in retrieval system. Therefore, the Eq. (2) can be modified for CLIR as follows:

\[
\Theta_{(l)} = \frac{tp_{(l)}}{tp_{(l)}+fn_{(l)}}
\]

(2)

where, \( tp \) represents true positives i.e. relevant documents in language ‘2’ from the retrieved document results

\( fn \) represents false positives i.e. relevant documents in language ‘2’ that are missed from the retrieved results

This work, used the standard recall 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 recall metric in evaluating performance of the MLIR systems.

The Eq. (2) is important with the stated reason in the above paragraph. Here ‘\( l \)’ specifies the language involved in the evaluation process.
\[ \Theta_{(2)} = \frac{tp_{(2)}}{tp_{(2)} + fn_{(2)}} \]  

(3)

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

\( fn \) represents false positives i.e. relevant documents in language \( '1' \) that are missed from the retrieved results

The above Eq. (3) is used to gives the recall value for the language \( '2' \). If two languages i.e. \( n = 2 \) are involved then CLIR systems performance is evaluated with \( \Theta_{(1)} \) and \( \Theta_{(2)} \).

\[ \partial R_{(1)} = \frac{\partial(1)}{\partial(2)} \]  

(4)

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

\[ \Theta_{(2)} = \frac{\partial(2)}{\partial(1)} \]  

(5)

\[ \partial U_{(1)} = \frac{\partial(1) - \partial(2)}{\partial(1)} \]  

(6)

\[ \partial U_{(2)} = \frac{\partial(2) - \partial(1)}{\partial(1)} \]  

(7)

If \( \partial U_{(1)} \) or \( \partial U_{(2)} \) gives positive value then it is literally upward performance/better performance otherwise it is literally downward performance/better performance. When the comparative values of recall between two languages reaches the average or not is checked using the Eqs. (6) and (7). If \( \partial U_{(1)} \) gives negative value then the relevant documents in language \( '1' \) are very less, that is, a wide enough search have to be done to retrieve most of the relevant documents in language \( '1' \) even though evaluation is required for more non-relevant documents. On the other hand, a narrow search has to be done then most of what we retrieve will be relevant, but there is chance to miss more relevant documents. In this case, low recall is achieved. High recall is tedious to achieve. If \( \partial U_{(2)} \) gives negative value then the necessary documents in language \( '2' \) are very less, that is, a wide enough search have to be done to retrieve most of the relevant documents in language \( '2' \) even though evaluation is required for more non-relevant documents. High recall means the almost all the actually relevant documents have been found i.e. true positives rate is very high.

4. Metrics for MLIR System Evaluation

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

\[ \Theta_{(i)} = \frac{tp_{(i)}}{tp_{(i)} + fn_{(i)}} \]  

(8)

where \( i = 1 \) to \( N \),

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

\( fn \) 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 \( \Theta_{(1)}, \Theta_{(2)}, ..., \Theta_{(i)} \) 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 \( '1' \) and in the place of 2 put \( '2' \) especially when comparing recall values among languages involved in retrieval process.

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

**Normalized performance of Recall:**

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

\[ \Theta_{(i)} = \frac{\partial(1)}{\partial(2)} \]  

(9)

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

\[ \Theta_{(3)} = \frac{\partial(1)}{\partial(2)} \]  

(10)

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

4. Experimentation and Result Analysis
There are variants presented in the literature for standard recall. Recall variants are: Recall(1000), R@R, R-precision, F-measure, etc. None of these exhibit performance in terms of individual language that is involved. Derivation of novel recall metric variants supposed to do a language wise performance evaluation. That is, $\Theta_{(1)}$ with respect to language ‘1’ and $\Theta_{(2)}$ with respect to language ‘2’. This kind of evaluation is essential where user needs 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 traditional metrics. This paper also concentrates on mapping the characteristics/attributes with standard recall metrics and the proposed recall metric variants. The following properties are mapped with recall are: Retrieval effectiveness, robustness, relevancy, efficiency and consistency.

The following properties are mapped with proposed recall metric variants apart from the attributes mapped with standard recall: 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 recall 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 recall 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, Alta vista, 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 recall 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 recall 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 Recall Metric variants (Google)**

<table>
<thead>
<tr>
<th>CLIR</th>
<th>$\Theta_{R_{(1)}}$</th>
<th>$\Theta_{R_{(2)}}$</th>
<th>$\partial \Theta_{(1)}$</th>
<th>$\partial \Theta_{(2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-G</td>
<td>0.56</td>
<td>0.1012</td>
<td>0.248</td>
<td>-0.078</td>
</tr>
<tr>
<td>S-I</td>
<td>0.133</td>
<td>0.45</td>
<td>0.066</td>
<td>0.15</td>
</tr>
<tr>
<td>G-S</td>
<td>0.162</td>
<td>0.222</td>
<td>-0.018</td>
<td>0.022</td>
</tr>
<tr>
<td>I-F</td>
<td>0.281</td>
<td>0.341</td>
<td>-0.0187</td>
<td>0.0213</td>
</tr>
<tr>
<td>D-G</td>
<td>0.435</td>
<td>0.115</td>
<td>0.155</td>
<td>0.155</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>CLIR</th>
<th>$\Theta_{R_{(1)}}$</th>
<th>$\Theta_{R_{(2)}}$</th>
<th>$\partial \Theta_{(1)}$</th>
<th>$\partial \Theta_{(2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-G</td>
<td>0.545</td>
<td>0.121</td>
<td>0.214</td>
<td>-0.07</td>
</tr>
<tr>
<td>S-I</td>
<td>0.245</td>
<td>0.71</td>
<td>-0.1</td>
<td>0.18</td>
</tr>
<tr>
<td>G-S</td>
<td>0.114</td>
<td>0.612</td>
<td>-0.85</td>
<td>0.26</td>
</tr>
<tr>
<td>I-F</td>
<td>0.75</td>
<td>0.217</td>
<td>0.257</td>
<td>-0.11</td>
</tr>
<tr>
<td>D-G</td>
<td>0.05</td>
<td>0.4</td>
<td>-0.05</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Table 3. Performance results of Recall Metric variants (Bing)**
Table 4. Performance results of Recall Metric variants (Ask)

<table>
<thead>
<tr>
<th>CLIR</th>
<th>Run</th>
<th>$\mathcal{R}<em>{R</em>{(1)}}$</th>
<th>$\mathcal{R}<em>{R</em>{(2)}}$</th>
<th>$\mathcal{U}<em>{U</em>{(1)}}$</th>
<th>$\mathcal{U}<em>{U</em>{(2)}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-G</td>
<td>0.675</td>
<td>0.2</td>
<td>0.225</td>
<td>-0.1</td>
</tr>
<tr>
<td></td>
<td>S-I</td>
<td>0.1</td>
<td>0.8</td>
<td>-0.1</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>G-S</td>
<td>0.45</td>
<td>0.44</td>
<td>0.15</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td>I-F</td>
<td>0.35</td>
<td>0.506</td>
<td>-0.04</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>D-G</td>
<td>0.2</td>
<td>0.36</td>
<td>-0.04</td>
<td>0.06</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>CLIR</th>
<th>Run</th>
<th>$\mathcal{R}<em>{R</em>{(1)}}$</th>
<th>$\mathcal{R}<em>{R</em>{(2)}}$</th>
<th>$\mathcal{U}<em>{U</em>{(1)}}$</th>
<th>$\mathcal{U}<em>{U</em>{(2)}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-G</td>
<td>0.52</td>
<td>0.311</td>
<td>0.083</td>
<td>-0.059</td>
</tr>
<tr>
<td></td>
<td>S-I</td>
<td>0.84</td>
<td>0.09</td>
<td>0.442</td>
<td>-0.099</td>
</tr>
<tr>
<td></td>
<td>G-S</td>
<td>0.34</td>
<td>0.432</td>
<td>-0.027</td>
<td>0.0324</td>
</tr>
<tr>
<td></td>
<td>I-F</td>
<td>0.082</td>
<td>0.94</td>
<td>-0.10</td>
<td>0.0578</td>
</tr>
<tr>
<td></td>
<td>D-G</td>
<td>0.243</td>
<td>0.456</td>
<td>-0.056</td>
<td>0.0863</td>
</tr>
</tbody>
</table>

Table 6. Normalized Recall performance

<table>
<thead>
<tr>
<th>MLIR normalized recall $\mathcal{R}<em>{N</em>{(i)}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Languages</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>S</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

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 recall measure’s language wise comparisons. The phase I results are described as follows: The performance of cross lingual run F-G is better than the other cross lingual runs. French queries are translated efficiently in Google search engine and because of this reason more relevant documents in German are recalled. So, $\mathcal{R}_{R_{(1)}}$ is high and $\mathcal{R}_{R_{(2)}}$ is low. When we measure $\mathcal{R}_{R_{(1)}}$ all the runs are good in performance except S-I and G-S. When measuring $\mathcal{R}_{R_{(2)}}$ S-I is better than the other runs. Italian retrieved documents are not better with the French retrieved documents in fig. 3. Average performance is given by I-F with $\mathcal{R}_{R_{(1)}}$ and G-S with $\mathcal{R}_{R_{(2)}}$. Poor performance is presented by the D-G cross lingual run. In both of the relative performance metrics the similar performance is shown by G-S and I-
F. In the fig. 4 almost similar documents are given by the I-F run when compared to F-G run of fig. 3. Very less number of documents is available in German Language as shown in fig. 5. Over all relative performance good with F-G and S-I cross lingual runs. On the other hand the metric $\Theta_U(1)$ is measured well for the language runs F-G this shows the better performance. The bitter performance is produced by all other language runs. S-I and D-G are measured well for the metric $\Theta_U(2)$ and that shows the better performance. Bitter performance is shown in fig 3 to 5 by the French Language. In fig. 5 the language run G-S is running negative side i.e. the number of relevant documents almost nil. The fig. 6 shows all the recall metrics are measured well when compared to other figures. In the same way, the next place is occupied by the fig. 7.

Phase II results are analyzed as follows: the above phase I result analysis is done with cross-lingual runs. When multiple languages are involved in the retrieval the performance is varied among the retrieval language documents. The normalized recall performance ($\Theta_N(i)$) is measured with five languages. Usually, in IR systems the range for measuring metrics is within the range 0 to 1. The fig. 8 shows that the language ‘F’ is better in performance rather than the other languages which are involved in MLIR retrieval. After that language ‘I’ is better. The language ‘F’ documents are reaching the value ‘1’. That is, in this language we will get 96% relevant documents and 90% of Italian documents. The performance is not repeated every time with same retrieval systems because the queries are different and the query languages using for translation is different. The language ‘G’ is having the average performance. That is, in almost all the retrieval systems ‘G’ relevant documents are marginal.

In general, few language’s relevant documents are high and other language’s relevant documents are low. This is because there may be ambiguities occurred in query formulation and translation. May be less number of documents available in ‘I’ and among them very few are relevant. Retrieval performance is varied between retrieval systems and the availability of various language documents is not equal. There are other reasons to say, these are few. From this, we conclude that retrieval system will be effective and efficient in terms of good document collection, good indexing system and finally good ranking model.

Figure 3. Performance evaluation of recall metric variants in Google

Figure 4. Performance evaluation of recall metric variants in Yahoo
Figure 5. Performance evaluation of recall metric variants in Alta vista

Figure 6. Performance evaluation of recall metric variants in Ask

Figure 7. Performance evaluation of recall metric variants in Bing
6. CONCLUSIONS

This paper presented the enhanced versions of the basic measure i.e. Recall metric variants. It also enumerated the advantages and disadvantages of the traditional measures. The new measures for CLIR and MLIR systems are derived and discussed. Because in these systems, the language wise comparisons are not easy with traditional IR recall 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 Recall metric.

High Recall is tedious to achieve. This effect is applicable when two or more languages 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 are involved. Because of language relatedness, importance of the language for the user when much language involved, tools of the languages (query or document translation) may be different, etc. So, the recall oriented analysis over the languages is very significant in the new era.

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

REFERENCES


[23] Olson, David L.; Delen, Dursun "Advanced Data Mining Techniques” Springer; 1 edition (February 1, 2008), page 138, ISBN 3540769161


