Asian Journal of Information Technology 13 (6): 346-355, 2014
ISSN: 1682-3915

A Metric to Assess the Performance of MLIR Systems

P. Sujatha and P. Dhavachelvan
Department of Computer Science, Pondicherry University, Puducherry, India

Abstract: Information retrieval plays a vital role in extraction of relevant information. Many researches have been working on for satisfying user needs, tough the problem arises when accessing multilingual information. This multilingual environment provides a platform where a query can be formed in one language and the result can be in the same language and/or different languages. Performance evaluation of information retrieval for monolingual environments, especially for English are developed and standardized from its inception. There is no specialized evaluation model available for evaluating the performance of multilingual environments or systems. The unavailability of MLIR domain specific standards is a challenging task. This study presented enhanced metric to assess the performance of MLIR Systems over its counterpart IR metric. This analysis shows that the performance of the enhanced metric is better than the conventional metric. And also, these metric can facilitate the researchers and developers to improve the quality of the MLIR Systems in the present and future scenarios.

Key words: Performance, metric, property, translation, protocol

INTRODUCTION

The goal of IR is to retrieve documents that most directly relevant to the request of users. With the speed of access and the large scale of the information sources available today, users often wish to reach beyond single information source in looking for relevant answers to their queries. MLIR Systems have also to address the problem of documents content representation and the problem of relevance evaluation. This evaluation is more difficult than in monolingual IR. It is indeed difficult to build a correspondence function with different languages for the documents and the query.

Traditional retrieval techniques creates an illusion that it is presenting all relevant documents to the user but it is not so, i.e., from the hit list only one fourth of the documents are useful according to the user need. Thus relevance varies from one user to another, for example for one user 20 documents may be relevant whereas for another user 40 documents may be relevant. Based on this binary relevance, all the documents in the resultant list are assessed and it is called continuous retrieval. An important consideration in evaluating MLIR Systems is the need of the user and the knowledge of languages since, a Multilingual System is most likely to be used in an interactive setting. It provides a means of measuring the quality of unranked information retrieval within a system. In this scheme, instead of simply employing the ranking produced by each SRE values, one can use the full potential of SRE values. This may be done by defining a measure that evaluates the utility or goodness of each SRE value. That is F-measure uses the full potential of all the SRE values and it is the measure of performance that takes into accounts both recall and precision.

LITERATURE REVIEW

The idea of using mathematical approaches for evaluating the performance of Information Systems has been initiated in the year of 1976 itself. Bookstein and Cooper (1976) presented a mathematical model of an information retrieval system thought to be general enough to serve as an abstract representation of most Documents and Reference Retrieval Systems. Savoy (1997) investigated several significance measures and echoed Van Rijsbergen's concerns using statistical methods. He proposed an Alternative Bootstrap Method, based on sampling from a set of query outcomes; it is not clear whether this approach could be applied with the small sets of queries for which the relevant judgments and it has not been used in practice to assess significance of these systems.

Li and King (1999) presented a new content-based retrieval approach using local MEFs extracted by the Principal Component Analysis (PCA) Method. When the number of features is large, it is important to select a set of MEFs to capture the majority of characteristics of an object and ignore the minor details. They used PCA to estimate the most expressive features and the advantage is to increase the retrieval precision through local feature selection.

Corresponding Author: P. Sujatha, Department of Computer Science, Pondicherry University, Puducherry, India

The main problem associated with this method is the parameter selection: unsuitable choices of the number of local MEFs for clusters will decrease the performance of the PCA retrieval approach.

Mandl (2008) presented an overview of the current activities of the major evaluation initiatives. Special attention is given to the current tracks and developments with the TREC, CLEF and NTCIR schemes. Researchers elaborated the basic activities and the history of the three major evaluation initiatives. Peters *et al.* (2001) presented the CLEF which promotes research into the development of truly multilingual systems capable of retrieving relevant information from the collections in many languages and in mixed media. It yielded good results only for cross lingual systems only and therefore it was possible to promote the same as a generic evaluation model for MLIR Systems.

Clough *et al.* (2008) discussed about the large-scale interactive evaluation of multilingual information access systems as a part of the CLEF campaign. In particular, the evaluation planned in 2008 which is based on the interaction with the content from Flickr, the popular online photo sharing service was described. Their proposed evaluation model seeks to reduce entry costs, stimulate user evaluation and encourage greater participation in the interactive track of CLEF. Yoshinaga *et al.* (1999) proposed new information retriever on the web which automatically classifies collected documents and retrieves multi-lingual information. In this research, unique properties were evaluated by a set of metrics as an extension of IR Systems but not as a domain specific model for MLIR Systems.

Yang and Lee (2008) specified that there are increasing needs in searching web pages of different languages using single query. In this research, a method based on GHSOM was proposed to discover the associations between different languages and it was applied to MLIR tasks. The experiments encouraged this approach to improve the MLIR performance but the evaluation scheme has been made as application-specific, not as a generic one.

Meng (2006) discussed about the traditional performance measures of information retrieval systems include precision and recall and their variants work well in closed-laboratory environments and also proved that they are not suitable for practical IR Systems such as web based search engines. This research presented many single-value measures to improve the precision-recall measures such as Expected Search Length (ESL), Average Search Length (ASL) and Rank Power. These metrics were applied to a set of real web retrieval data and compared their performance. They demonstrated that Rank Power is effective and easy to use as a single-value measure for performance of practical IR Systems but realized that it is not be suitable for the MLIR Systems.

Fujii and Ishikawa (2001) described a system for evaluating Japanese/English CLIR and MLIR Models, in terms of the retrieval accuracy and clustering effectiveness which are relied on the traditional metrics of IR Systems. Peters *et al.* (2001) aimed to demonstrate the importance of evaluation initiatives with respect to the system research and development. The main achievements of CLEF and the efforts that have been made to ensure that CLEF continues to meet the emerging needs of system developers and application communities were discussed. The discussions were on developers and application communities needs but not on the retrieval effectiveness which must be an important characteristic for a MLIR Systems.

Lin and Chen (2004) addressed a merging problem in the distributed MLIR Systems and several merging strategies have been addressed and also provided an unique evaluation scheme for the problems addressed. But no evaluation scheme has been provided for assessing the MLIR Systems and on the other hand, this research also increased the responsibility of the researchers in order to expand the scope of MLIR Evaluation Systems such that to accommodate the procedures for evaluating merging properties too.

Tune *et al.* (2007) discussed the design and implementation of dictionary-based CLIR System for indigenous language like Afaan Oromo, testing the performance of Oromo-English CLIR System at standard and internationally recognized evaluation forum like CLEF and demonstrating the feasibility CLIR application for nonwestern and resource scarce language like Afaan Oromo. Sethuramalingam and Varma (2008) described English to Hindi and Hindi to English CLIR Systems and the experiments were conducted using the FIRE-2008 dataset. Dictionary based approach was used for query translation and transliteration of named entities in the queries using a mapping-based, Compressed Word Format (CWF) algorithm. In both of the above cases, the overall system performance has been improved and they were validated using existing IR metrics set up and new metrics have been proposed in appropriate to the domains discussed.

In summary, all these researches are related to implementation of different IR/MLIR/CLIR Systems. Very few researches have concentrated on performance evaluation schemes based on MLIR evaluation initiatives. In the above presented survey most of the researches employed evaluation methodologies of IR Systems to evaluate MLIR Systems. This aspect is not appropriate to assess the performance of MLIR Systems and this issue

motivated us to propose a specific scheme of metric for assessing the performance of MLIR Systems. This metric will help the researchers and developers to improve the overall quality of the MLIR Systems in accordance to the present and future scenarios. Hence, the outcomes of this proposal will serve as a benchmark for MLIR Systems such that the researchers can make use of these metrics to evaluate their systems to confer the standards and benchmark performance claims. In addition to these, the ethnic findings of this research may lead to have a set of new research directions in MLIR paradigm.

MLIR SPECIFIC MEASUREMENT SCHEMES

Researchers have derived and evaluated MLIR specific scheme of metrics based on binary relevance and continuous retrieval. Basically, there are two important terminologies associated with the IR/MLIR domain: relevance and retrieval. Relevance has played a crucial role in MLIR research from 1950s which delineates the number of relevant documents from the retrieved documents. Retrieval delineates the number of documents retrieved in response to the query from total documents in the collection. Furthermore, relevance is decided by the user and retrieval of particular documents is decided by retrieval system. There are two types of relevance namely, Binary relevance and Ranked relevance. On the other hand, there are three types of retrievals namely, Binary retrieval, Ranked retrieval and Continuous retrieval. Using these notations, Stefano Mizzaro proposed four scenarios: Ranked relevance and Ranked retrieval, Binary Relevance and Ranked retrieval, Binary relevance and Binary retrieval and Binary relevance and Continuous retrieval (Mizzaro, 2001). In this study, researchers are going to expound the metrics related to binary relevance and continuous retrieval scheme.

Metrics based on binary and continuous retrieval measure: The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes. Traditional retrieval techniques creates an illusion that it is presenting all relevant documents to the user but it is not so, i.e., from the hit list, only one fourth of the documents are useful according to the user need.

Thus, relevance varies from one user to another for example for one user 20 documents may be relevant whereas for another user 40 documents may be relevant. Based on this binary relevance, all the documents in the resultant list are assessed and it is called continuous retrieval. An important consideration in evaluating MLIR Systems is the need of the user and the knowledge of languages, since a multilingual system is most likely to be used in an interactive setting. It provides a means of measuring the quality of unranked information retrieval within a system.

In this scheme, instead of simply employing the ranking produced by each SRE values, one can use the full potential of SRE values. This may be done by defining a measure that evaluates the utility or goodness of each SRE value. That is, F-measure uses the full potential of all the SRE values and it is the measure of performance that takes into accounts both recall and precision. Binary relevance and continuous retrieval measure are directly related to the system design issues.

Design issues

Scalability: The definition for scalability, according to Wiktionary is "the ability to support the required quality of service as the system load increases without changing the system." Thus, Scalable Information Systems are those systems that continue to sustain the required quality of service even under an increased load.

Retrieval performance: The measures require a collection of documents and a query. All common measures described here assume a ground truth notion of relevancy: every document is known to be either relevant or non-relevant to a particular query. Many different measures for evaluating the performance of Information Retrieval Systems have been proposed but most of the measures are Mono-Lingual.

Minimum error rate: IR System is said to be optimal only if it generate accurate results each time a query is processed. Many measure has be developed to calculate the errors rate occurring in IR Systems (e.g., precision at k, MAP, R-precision, etc.) (Mendenhall *et al*., 1990). Figure 1 shows the design issues and metric related. The metrics F-measure has been enhanced to support MLIR specific systems.

Performance evaluation of existing vs. enhanced metric

F-measure: F-measure was derived by Van Rijsbergen (1979) and it provides a way of combining recall and precision to get a measure which falls between recall and precision. F-measure is the performance measure that takes into accounts both recall and precision and it also defined as the harmonic mean of precision and recall. The F-measure can be defined as in the Eq. 1.

When compared to arithmetic mean, recall and precision need to be high for harmonic mean to be high.

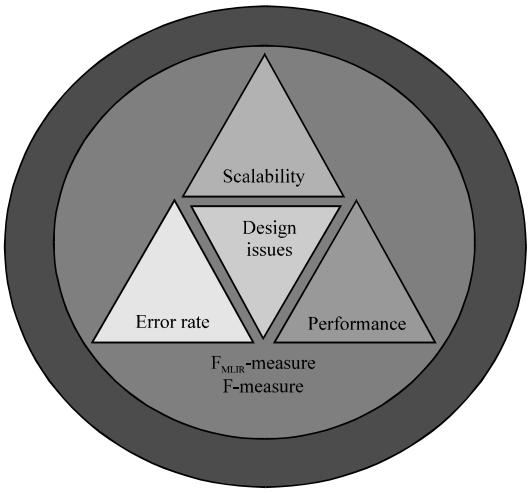

Fig. 1: Overview of design issues and related metrics

F-measure can be deduced as weighted average of the precision and recall, the best value to be '1' and worst value to be '0'. F-score is frequently used in the IR domain for measuring document classification, query classification and search performance. In terms of MLIR Systems when many languages are involved in the retrieval system, it is very intricate to identify precision and recall values in each language manually and obviously calculating F-measure is also not simple. Traditionally, this measure is used to predict the retrieval performance of the documents presented in one language. The advancements in internet and user needs lead to enhance this F-measure.

Existing metrics: F-measure:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}} \quad (1)$$

Where:
P = The precision
R = The recall

The advantage of this measure, it is single measure and popular for many other domains like medical and cross validation in machine learning, etc. The advantages which are carried out with precision and recall are also connected with this measure. On the other side, the drawback of this measure is calculating precision and recall is little time consuming task and knowing total number of relevant documents from the collection is a non-trivial task. Another drawback is when precision or recall is zero then there is no point of measuring effectiveness of retrieval system. In that case, calculating F-measure value becomes futile. In terms of MLIR Systems when many languages are involved in the retrieval system, it is very intricate to identify precision and recall values in each language manually and obviously calculating F-measure is also not simple. Traditionally, this measure is used to predict the retrieval performance of the documents presented in one language. The advancements in internet and user needs lead to enhance this F-measure. These viewpoints, suggested this research work to enhance the traditional F-measure to predict the retrieval effectiveness of MLIR Systems.

F_{MLIR}-enhanced measure: Retrieval of information on the MLIR Systems is entirely different from the retrieval in traditional IR Systems. Thus, the traditional evaluation methodology is not possible to measure different language queries and to assess them. In the past, measuring the effectiveness of retrieval systems was done using a few well known measures like precision, recall and F-measure (Jones, 1981). These measures can not be applied for these new retrieval scenarios. New or revised evaluation measures are required (Wang and Oard, 2006; Beg, 2005). The main aim of the MLIR Systems is to find relevant documents in numerous languages in response to a user query in any preferred languages. Hence, in a wider sense, it involves the process of determining documents that satisfies a user's information need. However, MLIR Systems and their evaluation have increased its importance. Therefore, the requirement to evaluate these systems more holistically cannot be overseen.

Though there are moderate number of evaluation initiatives are available, they adopt system centred approaches and failed to address the user centred approaches. This leads to greater challenging tasks in the near future to evaluate the MLIR Systems with reliable factors for use in the user centred approaches. As a result there is a need for both understand and present usable measures as well as methods of assessment for performing user-oriented issues. Holistic Evaluation Methods are needed to evaluate the new trends of MLIR Systems. Holistic in the sense which considers all the factors that are listed under user centred and system centred approaches. The aim of this research work is to achieve this goal in the evaluation of MLIR Systems. Thus, the need to maintain the balance is very much essential for overall goodness of the MLIR System. So, the quality and quantity will be improved when these combined factors in performance evaluation are applied. Therefore, the purpose of this research is to enhance the measures that will be usable and solve the issues involved in user centred approaches. This is based on the

fact that these are applied on the retrieved resultant lists of real life search engines. In this motive, traditional F-measure has been enhanced to achieve this goal of this research work. The enhanced method identifies, characterize and assess the traditional F-measure and modify it to adopt the user oriented aspects.

The F-measure is enhanced and used for performance evaluation of MLIR Systems because it is a popular approach for IR Systems' performance evaluation. The evaluation of these systems plays a crucial role to improve the quality of the systems. The foremost important way of evaluating the MLIR Systems is determining the retrieval effectiveness. In these MLIR Systems, the retrieval effectiveness evaluation measures deals with how effectively a given system can retrieve more relevant information and rank the relevant documents according to the user's information need:

$$F_{MLIR1} = \frac{2L_{r1}}{n+R_d} \quad (2)$$

Where:

r_1 = The number of documents retrieved and relevant in TL_1

n = The number of documents retrieved in TL_1

R_d = The total number of relevant documents in TL_1

The Enhanced Method identifies, characterize and assess the traditional F-measure and modify it to adopt the user oriented aspects. Equation 2 can be repeated for 'n' languages that are involved in the retrieval process. Precision, recall and F-measure are set-based measures (order of documents not taken into account) devised by researchers. In a continuous retrieval context, suitable sets of retrieved documents are obviously given by the top 'k' relevant documents. Retrieval effectiveness is measured manually but the quality of retrieval systems may not be measured effectively. User can easily identify which system is good in getting more relevant documents; in what languages the effectiveness is more than others.

Experimentation and result analysis: This study discusses four popular search engines such as Google, Bing, AltaVista and Yahoo. Google is selected because it is used by many web searchers and it is the largest vividly available search engine. Bing is a search engine that discovers and arranges the documents that the user needs, so the user can make faster and more informative decisions. The optimization technique of this system is similar to that of Google. AltaVista is a crawler-based search engine and it has efficacy to return diversified results at different times of the day. It supplies a free translation service, branded Babel Fish which automatically translates text between several languages. AltaVista gradually shed its portal features and refocused on search. Yet another search engine which is also popular and well-known in the Internet domain is Yahoo. The same set of queries and the relevance assessment methodology are used in these search engines.

English, French and Spanish language queries are used to retrieve the documents in these languages (Multilingual) from the specified search engines. The reason to choose these languages is that they are romance languages (many similarities) and Latin-based languages (Burr, 1998). Moreover, both (French and Spanish) languages have similar words and most of these query words used in the experiments have similar words. The dependent metrics for measuring F_{IR} is precision and recall. In this view point, Table 1 shows the measured values of precision and recall along with F_{IR} and F_{MLIR}.

The performance of Google is given in Table 1. For traditional F_{IR}, a monolingual run (E-E) is used and F_{MLIR} is calculated for the multilingual run (E-FSE). For each retrieval system, 20 queries were formed and submitted. For each query, the corresponding variables are measured and calculated for F_{IR} and F_{MLIR}.

From the Fig. 2, one can observe the performance differences between traditional and proposed measures. Very minimal differences are there between these two measures. For example, when a comparison is considered between F_{IR} and F_{MLIR} values of first query are 0.7998 and 0.8000, respectively, the difference between these two measures is in terms of 0.0001 values. The query numbers 2-4 have demonstrated the similar performance in both IR and MLIR Systems. The 50% of the queries show major differences between both F_{IR} and F_{MLIR}. The 35% of the

Table 1: Measuring F-measure in Google

Query No.	Run	Precision	Recall	F_{IR}	Run	F_{MLIR}
1	E-E	0.6400	0.9140	0.7528	E-FSE	0.7679
2	E-E	0.9000	0.9473	0.9230	E-FSE	0.9414
3	E-E	0.6100	0.8714	0.7176	E-FSE	0.7463
4	E-E	0.8200	0.9111	0.8631	E-FSE	0.8933
5	E-E	0.8500	0.9440	0.8945	E-FSE	0.9260
6	E-E	0.4400	0.5842	0.5019	E-FSE	0.5204
7	E-E	0.2000	0.4248	0.2719	E-FSE	0.2729
8	E-E	0.2100	0.6833	0.3212	E-FSE	0.3282
9	E-E	0.1900	0.3370	0.2429	E-FSE	0.2511
10	E-E	0.8100	0.2564	0.3895	E-FSE	0.3701
11	E-E	0.5500	0.7014	0.6165	E-FSE	0.6253
12	E-E	0.4400	0.1918	0.2671	E-FSE	0.2872
13	E-E	0.3500	0.6843	0.4631	E-FSE	0.4432
14	E-E	0.4400	0.4634	0.4513	E-FSE	0.4436
15	E-E	0.4500	0.2366	0.3101	E-FSE	0.3201
16	E-E	0.2600	0.2350	0.2468	E-FSE	0.2521
17	E-E	0.7000	0.5420	0.6109	E-FSE	0.6001
18	E-E	0.4900	0.7524	0.5934	E-FSE	0.6101
19	E-E	0.6100	0.4951	0.5465	E-FSE	0.5467
20	E-E	0.5400	0.1526	0.2379	E-FSE	0.2388

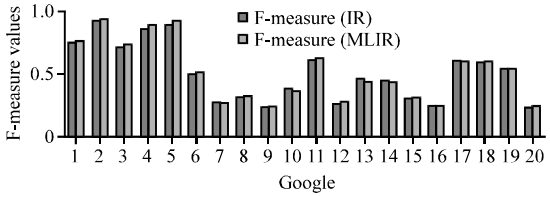

Fig. 2: Comparative evaluations of F_{IR} and F_{MLIR} using Google

Table 2: Measuring F-measure in Bing

Query No.	Run	Precision	Recall	F_{IR}	Run	F_{MLIR}
1	E-E	0.6800	0.9710	0.7998	E-FSE	0.8000
2	E-E	0.7500	0.9375	0.8333	E-FSE	0.8333
3	E-E	0.5800	0.8285	0.6823	E-FSE	0.6833
4	E-E	0.6800	0.9710	0.7998	E-FSE	0.8000
5	E-E	0.6700	0.9570	0.7881	E-FSE	0.7882
6	E-E	0.3300	0.7025	0.4490	E-FSE	0.5388
7	E-E	0.7700	0.6424	0.7004	E-FSE	0.6303
8	E-E	0.2500	0.7014	0.3686	E-FSE	0.3391
9	E-E	0.6400	0.1918	0.2951	E-FSE	0.3246
10	E-E	0.2400	0.6843	0.3553	E-FSE	0.4263
11	E-E	0.1400	0.4634	0.2150	E-FSE	0.2558
12	E-E	0.1900	0.2366	0.2107	E-FSE	0.2064
13	E-E	0.1800	0.2350	0.2038	E-FSE	0.1976
14	E-E	0.6800	0.5054	0.5798	E-FSE	0.5508
15	E-E	0.4300	0.5420	0.4795	E-FSE	0.4603
16	E-E	0.4600	0.7524	0.5709	E-FSE	0.5823
17	E-E	0.3400	0.4951	0.4031	E-FSE	0.4635
18	E-E	0.2300	0.1526	0.1834	E-FSE	0.2292
19	E-E	0.6600	0.5256	0.5851	E-FSE	0.5441
20	E-E	0.2300	0.9570	0.3708	E-FSE	0.3522

queries exemplify tiny differences among these two measures. The remaining 15% of the queries of both F_{IR} and F_{MLIR} have identical performances. This effect shows that MLIR performance is improved equally with the performance of IR counter parts.

Table 2 gives the measured values of F_{IR} and F_{MLIR} in Bing. Query number 4 shows high recall than the other 19 queries. Similarly query number 7 gives more precision than the other queries. As described earlier precision and recall should be high for F-measure to be high.

F_{IR} and F_{MLIR} values are high for the 2nd, 4th and 7th queries because their corresponding precision and recall values are high. Unlike Google, no other queries except the 2nd query have showed equal performance in F_{IR} and F_{MLIR} measures concern. Figure 3 shows the performance differences between the traditional and enhanced systems of 20 queries. The 80% of the queries show the major differences between F_{IR} and F_{MLIR}, 5% of the queries show equal performance and 15% of the queries demonstrate minute difference with the measured values. A very few queries demonstrated minimal differences between F_{IR} and F_{MLIR} values.

Table 3 depicts the experimental results of yet another retrieval systems' performance. For the monolingual run, the precision values of queries 1 and 4 are higher than the other precision values of the queries.

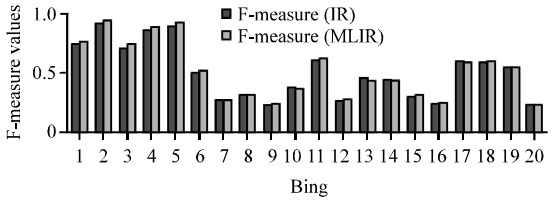

Fig. 3: Comparative evaluations of F_{IR} and F_{MLIR} using Bing

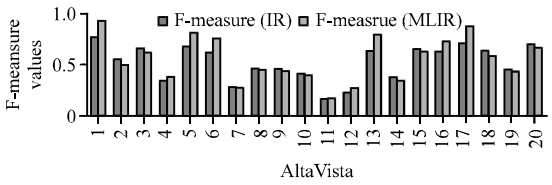

Fig. 4: Comparative evaluations of F_{IR} and F_{MLIR} using AltaVista

Table 3: Measuring F-measure in AltaVista

Query No.	Run	Precision	Recall	F_{IR}	Run	F_{MLIR}
1	E-E	0.8800	0.6887	0.7726	E-FSE	0.9271
2	E-E	0.7100	0.4637	0.5610	E-FSE	0.5049
3	E-E	0.6200	0.7322	0.6714	E-FSE	0.6176
4	E-E	0.8400	0.2219	0.3510	E-FSE	0.3861
5	E-E	0.6700	0.6958	0.6826	E-FSE	0.8191
6	E-E	0.7900	0.5256	0.6312	E-FSE	0.7511
7	E-E	0.3600	0.2372	0.2859	E-FSE	0.2801
8	E-E	0.3300	0.7679	0.4616	E-FSE	0.4477
9	E-E	0.6500	0.3633	0.4660	E-FSE	0.4427
10	E-E	0.3100	0.6531	0.4204	E-FSE	0.4035
11	E-E	0.2100	0.1472	0.1730	E-FSE	0.1764
12	E-E	0.2700	0.2127	0.2379	E-FSE	0.2735
13	E-E	0.6300	0.6366	0.6332	E-FSE	0.7915
14	E-E	0.6600	0.2630	0.3761	E-FSE	0.3497
15	E-E	0.7200	0.6137	0.6626	E-FSE	0.6294
16	E-E	0.5700	0.7261	0.6386	E-FSE	0.7343
17	E-E	0.8300	0.6147	0.7063	E-FSE	0.8828
18	E-E	0.7500	0.5553	0.6381	E-FSE	0.5934
19	E-E	0.5300	0.4026	0.4575	E-FSE	0.4346
20	E-E	0.9800	0.5440	0.6996	E-FSE	0.6646

When recall is concerned, only 10th query is out performed than the other recall values. Since, the precision and recall values of 11th query are poor, the F_{IR} and F_{MLIR} is also poor. But when compared to F_{IR}, F_{MLIR} performance is better in most of the queries. Highest F_{MLIR} value (0.9271) showed by the 1st query and in F_{IR} the highest value is just 0.7726 that was also exemplified by the 1st query. Only 20% of the queries are having highest values, 20% of the queries having the lowest values and other 60% of the queries performance is marginal. These above stipulated observations are perceived from the Fig. 4.

Earlier comparisons are made between the measured values of F_{IR} and F_{MLIR} using Google, Bing and AltaVista and here comparison between these measures using Yahoo is discussed. The poor performance is exemplified

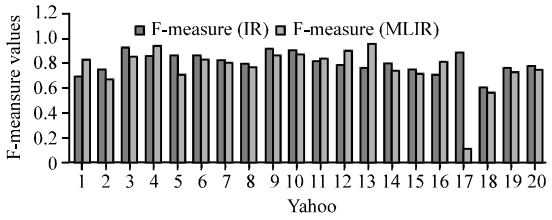

Fig. 5: Comparative evaluations of F_{IR} and F_{MLIR} using Yahoo

Fig. 6: Overall performance of F-measure in different IR Systems

Table 4: Measuring F-measure in Yahoo

Query No.	Run	Precision	Recall	F_{IR}	Run	F_{MLIR}
1	E-E	0.6300	0.7875	0.7000	E-FSE	0.8400
2	E-E	0.6800	0.8500	0.7555	E-FSE	0.6799
3	E-E	0.8900	0.9800	0.9328	E-FSE	0.8581
4	E-E	0.8200	0.9100	0.8626	E-FSE	0.9488
5	E-E	0.8300	0.9200	0.8726	E-FSE	0.7124
6	E-E	0.8300	0.9200	0.8726	E-FSE	0.8394
7	E-E	0.7900	0.8700	0.8280	E-FSE	0.8114
8	E-E	0.7400	0.8700	0.7997	E-FSE	0.7757
9	E-E	0.9000	0.9400	0.9195	E-FSE	0.8735
10	E-E	0.8900	0.9400	0.9143	E-FSE	0.8777
11	E-E	0.8900	0.7600	0.8198	E-FSE	0.8361
12	E-E	0.9600	0.6700	0.7892	E-FSE	0.9075
13	E-E	0.7000	0.8500	0.7677	E-FSE	0.9596
14	E-E	0.7200	0.9000	0.8000	E-FSE	0.7440
15	E-E	0.8700	0.6700	0.7570	E-FSE	0.7191
16	E-E	0.6300	0.8100	0.7087	E-FSE	0.8150
17	E-E	0.8800	0.9100	0.8947	E-FSE	0.1183
18	E-E	0.6700	0.5600	0.6100	E-FSE	0.5673
19	E-E	0.6900	0.8700	0.7696	E-FSE	0.7312
20	E-E	0.8700	0.7200	0.7879	E-FSE	0.7485

by the 1st query in this system for F_{IR} measure as shown in Table 4. The 60% of the queries of F_{MLIR} are better than the F_{IR} but least performance is demonstrated by query number 17 of F_{MLIR} which is very small value when compared to the F_{IR} value of 1st query (i.e., 0.7000).

On the other hand, best performance is produced by F_{MLIR}, i.e., 0.9596 than F_{IR}. These observations are exhibited from the Fig. 5. The remaining 40% of the queries of F_{MLIR} are not better when compared to the values of F_{IR}.

Figure 6 shows the comparison of four IR Systems by measuring the F_{IR}. Yahoo outperforms the other retrieval systems when measuring harmonic mean of precision and recall. For the 2nd query, F-measure of the Google outperforms the other retrieval systems and its

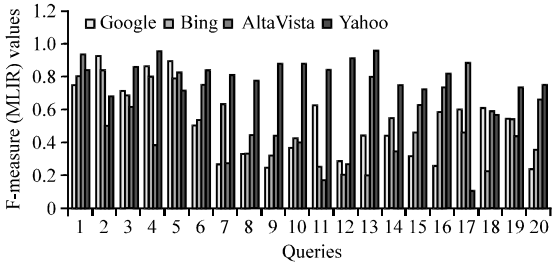

Fig. 7: Overall performance of F-measure in different MLIR Systems

value is more than any other queries of the Retrieval Systems. The next overall performance is produced by AltaVista. Google and Bing show equal performance except for the 2nd query. The most popular vividly used search engine Google's performance is good for the 18th query. When the overall performance is concerned, like F_{IR}, Yahoo outperforms the other retrieval systems. Google shows 50% better performance than the other retrieval systems like Bing and AltaVista. The next highest performance is demonstrated by AltaVista as like in F_{IR}. When compared to other retrieval systems Bing's performance is not even 35%. These experiments can be carried out for any number of languages that are provided by the search engines and can be done for any number of search engines that should have the capability to have the common language preferences. In the literature, research works that were presented have 50% difference between IR and MLIR. That is, MLIR performance was only half of the traditional IR Systems. Moreover, IR researcher in the trying to achieve equal performance curves with both IR and MLIR Systems. F_{MLIR} measure shows the equal performance with the IR measures in some of the cases. However, the F-measure values of MLIR Systems are more when compared to the F-measure values of IR Systems from the Fig. 6 and 7. Therefore, in this study, performance of F_{MLIR} crosses the performance of traditional IR Systems.

In this study, issues of detecting meaningful differences in F_{IR} and F_{MLIR} are discussed. Apart from this, the issues of detecting and measuring the similarities or differences between Retrieval Systems are also discussed without regard to any choice of performance measures that are presented in this dissertation. In this statistical analysis, the popular t-test and Wilcoxon signed rank tests are applied to know the significance of the two systems. The result analysis is discussed on four Retrieval Systems performance related to the F-measure but the statistical analysis of this study presents significance of F-measure for two retrieval systems. The hypothesis of these experiments is stated as follows:

Table 5: Independent t-test group statistics of Google (group statistics)

Groups	N	Mean±SE	t-value (sig.)
IR/MLIR			
F_{IR}	5	0.141310±0.0292253	-0.796 (0.002*)
F_{MLIR}	5	0.422050±0.0308532	-

Table 6: Independent t-test group statistics of Bing (group statistics)

Groups	N	Mean±SE	t-value (sig.)
IR/MLIR			
F_{IR}	5	0.141310±0.0280253	-9.894(0.009*)
F_{MLIR}	5	0.530330±0.0319818	-

Table 7: Wilcoxon signed rank test statistics of Google

Ranks	N	Mean rank	Z-value (sig.)
F_{IR}-F_{MLIR}			
Negative ranks	1	0.03	-2.791 (0.018*)
Positive ranks	4	2.00	-

Table 8: Wilcoxon signed rank test statistics of Bing

Ranks	N	Mean rank	Z-value (sig.)
F_{IR}-F_{MLIR}			
Negative	1	0.00	-2.761 (0.01*)
Positive	4	2.50	-

- Null hypothesis (H0): there is no difference between F_{IR} and F_{MLIR}
- Alternative hypothesis (HA): there is difference between F_{IR} and F_{MLIR}

Independent t-test: The independent t-test compares the mean values between two unrelated groups on the same continuous, dependent variable. The t-test procedure allows the testing of equality of variances (Levene's test) and the t-value for both equal and unequal-variance. For these tests the significance level (alpha) is set to 0.05 to either reject or accept the alternative hypothesis.

In case of Google, $p<0.05$ but in case of Bing there is no significance between two groups ($p>0.05$). The group statistics of the independent t-test are exemplified in Table 5 and 6 of Google and Bing Systems, respectively. Obviously, the mean values of the two groups may be similar because F_{IR} and F_{MLIR} of Bing are saying that they are equal. On the contrast, mean values of F_{IR} and F_{MLIR} of Google are different because F_{MLIR} performance is more than the F_{IR}.

Independent t-test group statistics are discussed and this test is one of the existing parametric tests. Table 6 states that F_{MLIR} is essential for evaluating the performance of the MLIR Systems because difference is there between traditional measure and enhanced measure.

Wilcoxon signed rank test: Wilcoxon signed rank test is a non-parametric test, i.e., it is used when data sets are not following normal distribution and it is used to test the median difference in paired data. Paired data means that the values in the two groups being compared are naturally linked and usually arise from individuals being measured more than once. Thus, to test F_{MLIR} metric performance in different cases, Wilcoxon test is performed. The null hypothesis is that the median difference between pairs of observations is zero. SPSS output of the Wilcoxon signed rank test is enumerated in Table 7. On performing the Wilcoxon signed rank test, it is noticed that there is statistical significance ($p>0.05$) between the MLIR and IR of F-measure.

In both the cases of this test, H0 is rejected as the $p<0.05$. Table 7 shows the significance value of two groups for Google (0.01) and Table 8 shows the significance value of two groups for Bing (0.018). Therefore, though the Wilcoxon signed rank test is a non-parametric test results of this test may not be taken into consideration because this may not be true for all the cases. For instance, in the experiments of the retrieval systems, F_{IR} and F_{MLIR} showed 55% of the queries are having same values with tiny differences. Despite this, this test shows significant results in both Google and Bing. Parametric tests are more appropriate than non-parametric tests, especially in the IR domain.

DISCUSSION

This study discussed the measure comes under the scheme of Binary relevance and continuous retrieval. For multilingual scenarios, a measure called F_{MLIR} to evaluate the performance of MLIR Systems has been developed. The F_{IR} is enhanced and used for measuring the performance of MLIR Systems because it is a popular and vividly used measure for predicting the performance of IR Systems. In the MLIR Systems, the evaluation plays a crucial role to improve the quality of the systems. The foremost important way of evaluating the MLIR Systems is determining retrieval effectiveness. Experiments are conducted on real-time search systems such as Google, Bing, AltaVista and Yahoo using document collections of English, Spanish and French languages. The result analysis shows that the enhanced measure is important in the multilingual scenarios especially in the evaluation process. Empirical result of the measured value outperforms the manually measured value and also outperforms the monolingual IR value. It shows very minimal differences between empirical and impractical values measured in each language. Except the value of Spanish language, measured values of other two languages are performing better than the empirical results. The raw aggregated measure is subject to statistical noise in the analysis.

In order to shape the outcomes of the research or to get scientific meanings about the findings, paired statistical tests are used. In all these paired statistical tests the F_{MLIR} demonstrates the $p<0.05$ except in one case. This reveals that there exists difference between traditional and enhanced measures. The enhanced measure F_{MLIR} outperforms F_{IR} the traditional measure. When statistically analysed the t-test shows the true significant over the F_{MLIR} than the Wilcoxon signed rank test.

CONCLUSION

MLIR Systems are very popular and attractive to the internet users and researchers. Evaluating these systems play a vital role in the field of information retrieval. A standard scheme is designed and described in this study along with metric related to it. This scheme is called as Binary Relevance and Continuous Retrieval and metrics comes under this scheme is a core metric (F_{IR}) has been enhanced to work with multilingual information scenarios. This study also shows the evaluation of the core metric with the enhanced metric (F_{MLIR}). This evaluation scheme will help the researchers and developers to improve the overall quality of the MLIR Systems. Hence, the outcomes of this study will serve as a benchmark for MLIR Systems such that the researchers can make use of these metrics to evaluate their systems. In future the other foresaid evaluation schemes of metrics will be focused.

ACKNOWLEDGEMENT

This research is a part of the Research Projects sponsored under the Major Research Project Scheme, UGC, India, Reference No.: F. No. 41-640/2012 (SR) dt. 16-07-2012. Researchers would like to express their thanks for the financial support offered by the Sponsored Agency.

REFERENCES

Beg, M.M., 2005. A subjective measure of web search quality. Inform. Sci., 169: 365-381.

Bookstein, A. and W. Cooper, 1976. A general mathematical model for information retrieval systems. Lib. Q., 46: 153-167.

Burr, E., 1998. Teaching romance linguistics with on-line French. Italian and Spanish Corpora LAUD, Series B: Applied and Interdisciplinary Papers.

Clough, P., J. Gonzalo, J. Karlgren, E. Barker, J. Artiles and V. Peinado, 2008. Large-scale interactive evaluation of multilingual information access systems: The iCLEF Flickr challenge. Proceedings of the 30th European Conference on Information Retrieval, March 30-April 3, 2008, Glasgow, UK., pp: 33-38.

Fujii, A. and T. Ishikawa, 2001. Evaluating multi-lingual information retrieval and clustering at ULIS. Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text retrieval and Text Summarization, March 2001, Tokyo, Japan, pp: 250-254.

Jones, K.S., 1981. The Cranfield Tests. In: Information Retrieval Experiment, Jones, K.S. (Ed.). Butterworth, London.

Li, X.Q. and I. King, 1999. Information retrieval using local linear PCA. Proceedings of the 6th International Conference on Neural Information Processing, Volume: 3, November 16-20, 1999, Perth, Australia, pp: 867-872.

Lin, W.C. and H.H. Chen, 2004. Merging results by predicted retrieval effectiveness. Proceedings of the 4th Workshop of the Cross-Language Evaluation Forum, August 21-22, 2003, Trondheim, Norway, pp: 202-209.

Mandl, T., 2008. Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance. Informatica, 32: 27-38.

Mendenhall, W., D.D. Wackerly and R.L. Scheaffer, 1990. Mathematical Statistics with Applications. PWS-Kent, Boston.

Meng, X., 2006. A comparative study of performance measures for information retrieval systems. Proceedings of the 3rd International Conference on Information Technology: New Generations, April 10-12, 2006, Las Vegas, NV., pp: 578-579.

Mizzaro, S., 2001. A new measure of retrieval effectiveness (Or: What's wrong with precision and recalls). Proceedings of the International Workshop on Information Retrieval, September 19-21, 2001, Finland, pp: 43-52.

Peters, C., M. Braschler, K. Choukri, J. Gonzalo and M. Kluck, 2001. The future of evaluation for cross-language information retrieval systems. Proceedings of the 2nd Workshop of the Cross-Language Evaluation Forum, September 3-4, 2001, Darmstadt, Germany.

Savoy, J., 1997. Statistical inference in retrieval effectiveness evaluation. Inform. Process. Manage., 33: 495-512.

Sethuramalingam, S. and V. Varma, 2008. IIIT Hyderabad's CLIR experiments for FIRE-2008. Proceedings of the 1st Workshop of Forum for Information Retrieval Evaluation, October 2008, Kolkata.

Tune, K.K., V. Varma and P. Pingali, 2007. Evaluation of Oromo-english cross-language information retrieval. Proceedings of the International Joint Conference on Artificial Intelligence, January 12, 2007, Hyderabad, India.

Van Rijsbergen, C.J., 1979. Information Retrieval. Butterworths, London, UK.

Wang, J. and D.W. Oard, 2006. Combining bidirectional translation and synonymy for cross-language information retrieval. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 6-11, 2006, Seattle, pp: 202-209.

Yang, H.C. and C.H. Lee, 2008. Multilingual information retrieval using GHSOM. Proceedings of the 8th International Conference on Intelligent Systems Design and Applications, Volume: 1, November 26-28, 2008, Kaohsiung, pp: 225-228.

Yoshinaga, K., T. Terano and N. Zhong, 1999. Multi-lingual intelligent information retriever with automated ontology generator. Proceedings of the 3rd International Conference on Knowledge-Based Intelligent Information Engineering Systems, December 1999, Adelaide, SA., pp: 62-65.