Precision at K in Multilingual Information Retrieval

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

ABSTRACT
Information Retrieval (IR) is used to store and represent the knowledge and the retrieval of information relevant for a special user query. Multilingual Information Retrieval (MLIR) system helps the users to pose the query in one language and retrieve the documents in more than one language. One of the basic performance measures of IR systems is Precision. While this measure work well in monolingual web retrieval, not suitable for CLIR (Cross-lingual Information Retrieval) or MLIR where two or more languages are involved. This paper proposed a metric which measures Precision at K which is the proportion of relevant documents in the first K positions when more than one document languages involved in the retrieval system i.e. MLIR. Experimental results demonstrate that the proposed metric is effective in systems where more than one document languages involved in the retrieval.

General Terms
Information retrieval, cross lingual, Recall

Keywords
Multilingual, Precision, Recall, Retrieval effectiveness

1. INTRODUCTION
The ever increasing volume of information available in our daily lives is creating increasing challenges for document retrieval technologies [13]. MLIR system [1] helps the users to pose the query in one language and retrieve the documents in more than one language. A set of documents that are written in different languages is called a multilingual data collection. Two types of multilingual data collection are available in the literature. The first one contains several monolingual document collections. The second one consists of multilingual-documents. A multilingual-document is written in more than two languages. Some multilingual-documents have a major language, i.e. most part of the document is written in the same language [11]. Our work has used first type of data collection to measure the proposed metric. The MLIR techniques are: An approach for exploiting the Web as the multilingual corpus source for translating unknown query terms have been proposed by [2]. Many user queries contain terms not found in an ordinary translation dictionary. A novel technique is there to mine bilingual search-result pages obtained from a Web search engine for helping the translation of unknown query terms. This approach can also be used for improving a domain-specific bilingual lexicon. The same approach can also promote multilingual access in a digital library. In [3] the feasibility of employing various cross-lingual information retrieval techniques for developing and evaluating a bilingual Web portal was investigated. There are five major components in this approach, namely, Web spider and indexer, pre-translation query expansion, query translation, post-translation query expansion, and document retrieval. The query translation is based on a dictionary-based approach that includes phrasal translation, co-occurrence analysis, and pre- and post-translation query expansion. These techniques have been applied to the development of a multilingual Web portal, ECBizPort, which is an English–Chinese Web portal for business intelligence in the information technology domain.

The problem of MLIR is essentially one of machine translation on a very small scale. There are two documents approaches to this problem [14]. One is the dictionary translation using machine readable multi-lingual dictionaries and another is automatic extraction of possible translation equivalents by statistical analysis of parallel or comparable corpora. There is a serious problem of these MLIR systems is how users can estimate the relevance of retrieved documents that are represented in multiple languages and how they can choose the most relevant documents for computer or human translation.

Broadly used "Precision" and "Recall" are two measures of IR success [15], both measures based on the concept of relevance. Precision is defined as, "the ratio of relevant items retrieved to all items retrieved, or the probability given that an item is retrieved it will be relevant" [4]. A more common type of precision variants, widely used in the research community is *average precision [AP]*. This family of measures reflects the recognition that precision varies, generally falls, as recall increases. This variation can be articulated directly as a graph of precision vs. recall. This kind of curve is used to assess the different IR algorithms, or across different document collection of the same algorithm. Another variant of Precision is *Non-interpolated average precision*" corresponds to the area under superlative (non-interpolated) recall/precision curve." [5] This metric is measured by "computing the precision after every retrieved relevant document and then averaging these precisions over the total number of retrieved relevant documents" for a given query. There will be a different average precision, in general, for each query. These averages can then be themselves averaged over all the queries and it can be leads to yet another variant of precision called *Mean Average Precision (MAP*).

Measuring precision is pretty easy; if a set of users or judges agree on the relevance or non-relevance of each of the retrieved documents, then calculating the precision is straightforward. Of course, this assumes that the set of retrieved documents is of manageable size, as it must be if it is to be of value to the user. If the retrieved documents are ranked, one can always reduce the size of the retrieved set by setting the threshold higher, e.g., only look at the top 100 or the top 20 is called *P@*k which is another variant of precision. In this paper concentration is given on this variant of precision.

The paper is organized as follows: the various existing metrics of IR are discussed in section II, derivation of new metric and its importance in MLIR domain is stated in section III, the section IV gives the experimental results and finally section V concludes the paper.

2. RELATED WORK

In the literature, there are few papers regarding novel metrics for IR. They are: A novel evaluation metric PRES have been devised for recall-oriented patent retrieval task [6]. 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 rank-based metrics allowed the researchers to maximize different metrics for the same training data. [7] There are metrics which are optimized based on some smooth approximation with gradient descent; they are Normalized Discounted Cumulative Gain (NDCG) and AP. In [8], 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.

The recent research has suggested an evaluating IR systems based on user behaviour. The effectiveness of IR is usually evaluated using NDCG, MAP and P@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 [9], 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 [10]. 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.

P@K which is the proportion of relevant documents in the first K positions and is given below:

$$P@K = \frac{1}{k}\sum_{i=1}^{m} l_i \, 1(r(i) \le k) \qquad (1)$$

Where 1 is the indicator function: 1(A) = 1 is A is true, 0 otherwise [12].

Till now, to our knowledge, there are no metrics derived especially for MLIR systems. Traditional IR measures are used to measure the performance of these systems. As the first attempt, we have developed a new metric for MLIR systems to check the retrieval effectiveness when multiple languages are involved in retrieval process.

3. THE $P@K_{LDJ}$ METRIC

The traditional measure, precision takes all retrieved documents into account. It can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This metric is not effective for MLIR systems because the document collection consists of more than one language. The equation (1) need to be enhanced in order to check the performance of different languages at some K positions. For example in the given document collection 3 languages are involved. They are Telugu, Tamil and English. When a query is submitted to the MLIR system, *P@K* can be measured and the result would be combined as like monolingual IR. Language wise performance is not measured at K positions. Precision at top K documents of MLIR is measured using equation (2). This proposed metric is useful for understanding the retrieval effectiveness of multiple document languages.

$$P@k_{LD_j} = \frac{\left[\sum_{j=1}^{n}\left(\frac{starting\ position\ of\ LD_j}{ending\ position\ of\ LD_j}\right)*R\right]}{K} \qquad (2)$$

Where $R = LD_J \sum_{i=1}^{K} r_i$

Where:

'n' is the number of languages involved in the retrieval system,

LD_j is the document in j^{th} language

R is the number of relevant documents in '*n*' languages

r is 1 if the document in j^{th} language is relevant 0 otherwise *K* is the position of the ranked list

All the relevant documents in each language are counted and the corresponding starting and ending positions at the cut-off ranked position is calculated.

4. EXPERIMENTAL RESULTS

We have done the experiments on Google search engine. The query language is English and the document languages are French (F), German (G) and Hindi (H). For a query the $P@K_{LDj}$, is measured and demonstrated in Table I for 10 values i.e. k varies from 5 to 50. Table II is demonstrated the three monolingual runs E-E, E-H and E-F using the traditional metric P@K.

Table 1. $P@K_{Ldj}$ using multilingual run

Metric	E-FGH
P@5	0.2
P@10	0.3714
P@15	0.1524
P@20	0.1557
P@25	0.1252
P@30	0.1049
P@35	0.0902
P@40	*0.0791*
P@45	0.0706
P@50	0.0636

The traditional *P@K* measure relatively decreased when number of retrieved relevant documents is increased as depicted in Fig. 1. But the proposed metric value is varies as the number of retrieved relevant document in three languages increased for each cut-off value as illustrated in Fig. 2. By this analysis we can see the effectiveness and usefulness of the proposed metric. Traditional IR measure is very much suitable for monolingual IR and the proposed metric is adequate for measuring the MLIR retrieval effectiveness. The computed measure allows us to know the retrieval effectiveness when multiple languages are involved.

Table 2. Traditional measure, P@*K* using three monolingual runs

Metric	E-E	E-H	E-F
P@5	0.8	0.2	0.6
P@10	0.6	0.1	0.5
P@15	0.466	0.133	0.533
P@20	0.5	0.1	0.4
P@25	0.52	0.16	0.44
P@30	0.5	0.1333	0.4
P@35	0.4857	0.1142	0.3714
P@40	0.525	0.125	0.325
P@45	0.4666	0.111	0.3111
P@50	0.44	0.1	0.34

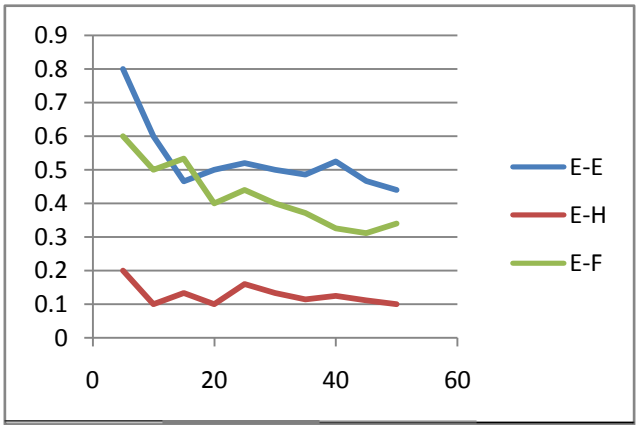

Fig 1: Performance of three monolingual runs

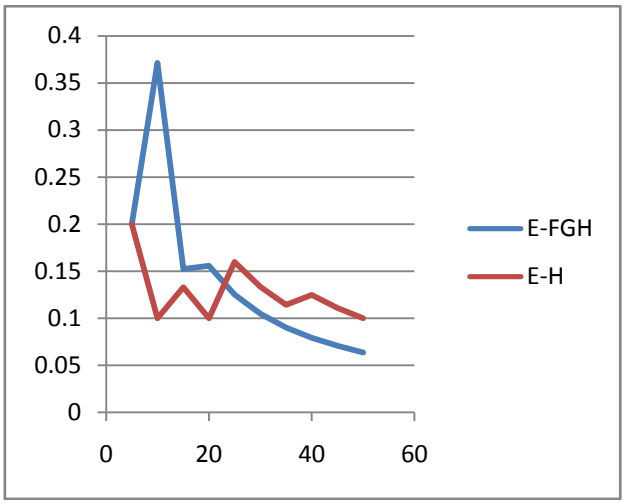

Fig 2: Performance of Monolingual run vs. Multilingual run

The same performance is not carried with three monolingual runs when compared to single multilingual runs. The figure 3 depicts the retrieval performance of two monolingual runs E-H, E-F and a multilingual run E-FGH. E-F runs behave well at all the cut-off values k of P because English and French language tools are vividly available at the same time both have similar preprocessing linguistics elements. That is the reason E-F run performance is better that E-H. On the other side multilingual run is better for P@5,10,15 and 20 than the E-H monolingual run. Particularly P@10 of E-FGH is double the performance of the monolingual run. Thus measuring $P@K_{LDj}$ is important when more than two languages are involved in the retrieval process.

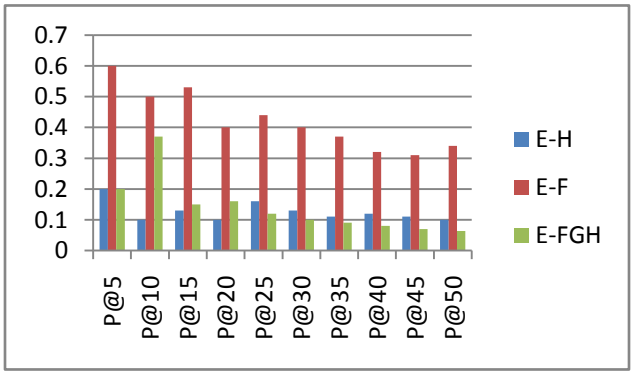

Fig:3 Performance of $P@K_{Ldj}$ of monolingual vs. MLIR

5. CONCLUSION

This paper discussed the new measure $P@k_{LDj}$, which measures the top K relevant documents in '*n*' number of languages. The traditional measure P@K is enough to measures the performance of monolingual information retrieval effectiveness since its value is always decreased as the number of relevant document are increased. The proposed measure is adequate to know the retrieval effectiveness of MLIR systems because for some cases the performance is not decreased (varied) as the number of relevant documents in various languages increased. Experiments were conducted using Google search engine by considering three languages. Result analysis demonstrated that the run E-FGH is better than one of the monolingual runs E-H.

6. REFERENCES

[1] Oard, D.W. and Dorr, B.J., 1996. A Survey of Multilingual Text Retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

[2] Wang, J.-H., Teng, J.-W., Lu, W.-H., & Chien, L.-F. (2006). Exploiting the Web as the multilingual corpus for unknown query translation. Journal of the American Society for Information Science and Technology, 57, 660–670.

[3] Qin, J., Zhou, Y., Chau, M., & Chen, H. (2006). Multilingual Web retrieval: An experiment in English–Chinese business intelligence. Journal of the American Society for Information Science and Technology, 57, 671–683.

[4] Saracevic, T. Evaluation of Evaluation in Information Retrieval, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 137-146, 1995.

[5] Harman, D. Overview of the Second Text Retrieval Conference (TREC-2), Text Retrieval Conference-2, Gaithersburg, MD, National Institute of Standards and Technology Special Publication 500-215, March, 1994.

[6] Walid Magdy, Gareth J. F. Jones: PRES: a score metric for evaluating recall-oriented information retrieval applications. Dublin City University, 33rd Annual ACM SIGIR Conference 2010, 611-618.

[7] Donald A. Metzler, W. Bruce Croft, and Andre McCallum: Direct Maximization of Rank-Based Metrics for Information Retrieval. July 16, Published at CIIR 2005.

[8] Olivier Chapelle, Mingrui Wu: Gradient descent optimization of smoothed information retrieval metrics. Learning To Rank For Information Retrieval, Volume 13, Number 3, June 2010, 216-235.

[9] 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.

[10] 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.

[11] W. C. Lin and H. H. Chen. Merging mechanisms in multilingual information retrieval. Lecture notes in computer science, pages 175–186, 2003.

[12] Chapelle, O., Wu, M.: Gradient descent optimization of smoothed information retrieval metrics. Information Retrieval Journal (To appear) (2010)

[13] W.Li, D.Ganguly and G.J.F.Jones,: Enhanced Information Retrieval Using Domain-Specific Recommender Models, In Proceedings of the 3rd International Conference on the Theory of Information Retrieval (ICTIR'11), Bertinoro, Italy, September 2011

[14] Mandl, Thomas; Womser-Hacker, Christa; Ferro, Nicola; Di Nunzio, Giorgio (2008).: How Robust are Multilingual Information Retrieval Systems? In: Proceedings A CM Symposium on Applied Computing (SAC) Fortaleza, Brazil. pp. 1132-1136.

[15] Buckley, Chris; Voorhees, Ellen (2005): Retrieval System Evaluation. In: TREC: Experiment and Evaluation in Information Retrieval. Cambridge & London: MIT Press. pp. 53-75.