A New Metric for Patent Retrieval Evaluation

Walid Magdy and Gareth J.F. Jones
Centre for Next Generation Localization
School of Computing
Dublin City University
Dublin 9, Ireland
{wmagdy, gjones}@computing.dcu.ie

ABSTRACT

Patent retrieval is generally considered to be a recall-oriented information retrieval task that is growing in importance. Despite this fact, precision based scores such as mean average precision (MAP) remain the primary evaluation measures for patent retrieval. Our study examines different evaluation measures for the recall-oriented patent retrieval task and shows the limitations of the current scores in comparing different IR systems for this task. We introduce PRES, a novel evaluation metric for this type of application taking account of recall and user search effort. The behaviour of PRES is demonstrated on 48 runs from the CLEF-IP 2009 patent retrieval track. A full analysis of the performance of PRES shows its suitability for measuring the retrieval effectiveness of systems from a recall focused perspective taking into account the expected search effort of patent searchers.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.4 Systems and software – *performance evaluation*.

General Terms

Measurement, Performance, Experimentation.

Keywords

PRES; Patent Retrieval; Evaluation Metric

1. INTRODUCTION

Interest in patent retrieval research has had a considerable growth in the recent years. Reflecting this, patent retrieval has been introduced as a task at two of the major information retrieval (IR) evaluation campaigns (NTCIR and CLEF) in 2003 and 2009 respectively. The aim is to encourage researchers into identifying the best IR methods for achieving the highest retrieval effectiveness for patent search. Patent retrieval is usually identified as a recall-oriented retrieval task, where the objective is to find all relevant documents [7]. For precision focused IR tasks, where one or two of the relevant documents are often sufficient for achieving user satisfaction and hence the objective is to find relevant documents as soon as possible, whereas for patent retrieval the objective usually aims to find all relevant documents even if more effort will be exerted by the user. Despite this fact, MAP is the most commonly used metric for evaluating patent retrieval.

Viewing patent retrieval as simply a recall-oriented task is actually rather simplistic. In practice the time and expense of patent searchers is limited, and thus an evaluation metric should take account not only recall, but the effort expended to achieve a given level of recall.

In this paper, we describe a study to analyze the behaviour of current evaluation metrics when applied to the patent retrieval task. The results of this analysis are used to motivate the proposal of a novel evaluation metric which combines recall with the quality of ranking the retrieved relevant results. Experimental evaluation demonstrates that the new score has high effectiveness for evaluation of patent retrieval with ranked output. The study is performed on the CLEF-IP 2009 patent retrieval task [16]. Forty-eight submitted runs in CLEF-IP 2009 task are used to compare the performance of this novel metric and the existing measures. The aim of the CLEF-IP track is to automatically find prior art citations for patents. The topics for this task are patents filed in the period after 2000, and the searched collection contains about one million patents filed in the period from 1985 to 2000 [16]. The objective is to use some text from each patent topic to automatically retrieve all cited patents found in the collection. These citations are originally identified by the patent applicant or the patent office.

The remainder of the paper is organized as follows; Section 2 surveys background on patent retrieval and IR evaluation scores; Section 3 explores the effectiveness of the current IR evaluation scores for measuring system performance for recall-oriented IR applications Section 4 explains normalized recall, which is one of the classic IR evaluation scores used later to develop our new PRES evaluation metric, Section 5 formally introduces PRES; Section 6 explores the behaviour of PRES by use of illustrative examples and by testing on the 48 CLEF-IP 2009 runs, and finally Section 7 concludes the paper with suggestions for possible future research directions.

2. BACKGROUND

2.1. Patent Retrieval

Evaluation of patent retrieval was proposed in NTCIR-2 in 2001 [13]. Since then patent retrieval has featured as a fixed track in all NTCIR[1] campaigns. Similar tasks around patent retrieval were introduced to CLEF[2] in 2009 carrying the name of CLEF-IP (CLEF Intellectual Property) [16]. This task has been of interest to IR researchers since its introduction due to the challenging nature

1st International Workshop on Advances in Patent Information Retrieval (AsPIRe'10), March 28, 2010, Milton Keynes.

[1] http://www.nii.ac.jp/

[2] http://www.clef-campaign.org/

of patents itself [13, 16]. Various tasks have been created around patents; some are related to IR and others such as patent mining and patent classification.

The IR tasks at NTCIR and CLEF related to patent retrieval are as follows

2.1.1 Ad-hoc search

A number of topics are used to search a patent collection with the objective of retrieving a ranked list of patents that are relevant to this topic [10]

2.1.2 Invalidity search

The claims of a patent are considered as the topics, and the objective is to search for all relevant documents (patents and others) to find whether the claim is novel or not [7]. All relevant documents are needed, since missing only one document can lead to later invalidation of the claim or the patent itself.

2.1.3 Passage search

The same as invalidity search, but because patents are usually long, the task focuses on indicating the important fragments in the relevant documents [8].

2.1.4 Prior-art search

In this task, the full patent is considered as the topic and the objective is to find all relevant patents that can invalidate the novelty of the current patent, or at least patents that have common parts to the current patent [16].

2.2. Evaluation Metrics

While many evaluation metrics have been proposed by ad hoc type IR tasks, by far the most popular in general used is MAP [3]. The standard scenario for use of MAP in IR evaluation is to assume the presence of a collection of document representative of a search task and set of test topics (user queries) for the task along with associated manual relevance data for each topic. The relevance data for each topic is assumed to be a sufficient proportion of the documents from the collection that are actually relevant to that topic. "Sufficient" here relating to the fact that the actual number of relevant documents each topic is unknown without manual assessment of the complete document collection for each topic. Several techniques are available for determining sufficient relevant documents for each topic [4, 11, 17]. As its name implies, MAP is precision metric, which emphasizes returning more relevant documents earlier. The impact on MAP of locating relevant documents later in the search of a ranked list is very weak, even if very many such documents have been retrieved. Thus while MAP gives a good and intuitive means of comparing systems for IR tasks emphasising precision, it will often not given a meaningful interpretation for recall focused tasks. Some other IR evaluation metrics are found to be more representative than MAP for others types of IR task. For example, Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG) are used for IR applications such as question answering and web search respectively [5, 18]. MRR measures performance when looking for one specific "known item" in the document collection [2]. Mean reciprocal rank is simply the inverse of the rank of the relevant document in the retrieved list. 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 the less relevant. Additional IR evaluation scores have been introduced with the advent of new IR applications such as mean average generalized precision (MAgP) for structured documents retrieval [1, 12] and GMAP which is the same as MAP but using geometric mean instead of the arithmetic mean, GMAP is used in the Robust Track at TREC [19].

Similar to MAP, these IR evaluation metrics focus on measuring the effectiveness at retrieving relevant documents earlier rather than on the system recall. While this is sufficient and reasonable for precision focused tasks where one or two relevant documents may be sufficient to satisfy the user, it is not suitable for tasks where the objective is to find "All" or at least significant proportion of relevant documents, and in particular if the objective is to find all relevant documents with minimum effort for the user. In this kind of application, the user is willing to exert much effort to go deeper in the list in order to find as many relevant documents as possible. For example, in patent retrieval, the design of the patent test collection assumes that filed patents examined by the patent office for novelty, are the training and test collections, and that the patent citations, which are mostly added by the patent office, are considered as the relevance assessment [7, 16, 9]. The recall of the relevant documents in the relevance assessment can be considered to be almost 100%, as much effort, time, and money are spent to identify these relevant cited documents, especially for issued patents which take years to be searched for novelty. Furthermore, all citations are for related technologies that do not invalidate the novelty of the patent, or otherwise the patent will not be issued.

For a recall-oriented IR application such as patent retrieval the maximum number of documents to be checked by the user is also very important, since it has a direct impact on the cost of user effort and on recall. This concern was the reason behind using recall along with MAP in evaluating similar IR tasks [16, 20]. The maximum number to be checked by the user is completely overlooked by most of the metrics considered so far, and is variable in measures such as the f-score [15]. The f-score combines recall with precision, and has been used for legal IR [14]; although this score carries recall in its formula, it has the problem that the number of documents to be retrieved is not fixed, which is usually a practical concern of patent officers.

3. IR EVALUATION SCORES FOR PATENT RETRIEVAL TASK

The simplest solution to measuring performance in a recall focused IR task is of course simply to evaluate the recall. However, as noted in the previous section, the problem of doing this is that it fails to reflect how early a system retrieves the relevant documents and thus the user effort involved. Although recall is the objective for such applications, the score should be able to distinguish between systems that retrieve relevant documents earlier than those that retrieve them later. To overcome this problem f-score can be used, but at a fixed number of retrieved documents. However the same problem will arise, as applying it after retrieving *N*-documents; for two systems that retrieved the same number of relevant documents, the f-score will be the same. F-score is designed for classification tasks, but for recall-oriented IR applications, the problem is viewed as a ranking problem with a cut-off for a maximum number of documents to be checked N_{max}.

A possible proposal for using the f-score is to calculate it as a combination between the recall and the average precision (*AP*) instead of using the absolute precision (equation 1). Such a modified f-score will reflect the system recall in addition to its average precision. However, while this captures the recall, it will have the same disadvantages for recall focused tasks with respect to *AP* which were noted earlier.

$$F'_{\beta} = \frac{(1+\beta^2)\cdot(AP \cdot R)}{\beta^2 \cdot AP + R} \qquad \textbf{(1)}$$

where, ***AP***: Average precision of a topic
R: recall at a given number of retrieved documents
β: weight of recall to precision

Table 1 shows an illustrative example on how different metrics perform with four different IR systems when searching a collection for a single query. In this case it is known that there are four relevant documents, and it is assumed that the user is willing to check the top 100 retrieved documents by each system.

Table1. Performance of different scores with different IR systems (Average precision, recall@100, f-score, modified f-score with different weights to recall)

	Ranks of rel. docs	***AP***	***Recall***	***F_1***	***F'_1***	***F'_4***
System 1	{1}	0.25	0.25	0.0192	0.25	0.25
System 2	{50, 51, 53, 54}	0.0481	1	0.0769	0.0917	0.462
System 3	{1, 2, 3, 4}	1	1	0.0769	1	1
System 4	{1, 98, 99, 100}	0.2727	1	0.0769	0.429	0.864

In Table 1, system 3 is the prefect with all relevant documents retrieved at the top ranks. System 1 has the lowest recall one, while system 2 has a moderate performance retrieving all relevant documents in the middle of the ranked list, System 4 has fair performance since it achieves 100% recall, but only after checking the full list of 100 top results. It can be seen that it achieves partially good performance by retrieving a relevant document in the first rank.

From the table it can be seen that *AP* for system 1 is much higher than for system 2, which is unfair, since system 2 has been able to retrieve all relevant documents in the middle of the list, which the user would be willing to check for, but system 1 has failed to retrieve more than one relevant document in the full list. The same situation arises when comparing system 4 to system 2, even though both systems have been able to retrieve the full list of relevant documents, system 2 has done so at much higher ranks than system 4.

The recall and F_1 scores fail to differentiate between systems 2, 3, and 4, even though these systems have very different behaviour.

F'₁ does not focus on the recall, which is the objective of recall-oriented applications. To emphasize recall a modified f-score, *F'₄* was tried giving recall four times the weight of the average precision. Initial inspection suggests that *F'₄* looks to be a good representation of the system performance, however on deeper analysis, it can be seen that system 4 is evaluated as nearly twice as good as system 2, even though while it retrieves a relevant document at rank 1 no further relevant documents are found until the end of the list and that while system 2 failed to return any relevant documents among the first half of the list, all relevant documents are retrieved by rank 54. For two systems such as 2 and 4 for a recall-oriented task with users willing to check the first 100 documents, system 2 will give more confidence to the user that there is a little chance of finding further relevant documents after rank 100, but system 4 will not give the user the same confidence, since the presence of low ranked relevant may suggest that further ones are likely to be present. Hence, *F'₄* fails to evaluate system 2 and system 4 in a fair way from the prospective of a recall-oriented application in practical usage.

4. NORMALIZED RECALL (R_{norm})

One of the proposed IR evaluation metrics that has never found its way into wide usage is normalized recall (R_{norm}) [15], shown in Equation 2. This measures the effectiveness in ranking documents relative to the best and worst ranking cases, where the best ranking case is retrieval of all relevant documents at the top of the list, and the worst is retrieving them only after retrieving the full collection. Figure 1 shows an illustrative graph of how to calculate R_{norm}, where R_{norm} is the area between the actual and worst cases divided by the area between the best and worst cases.

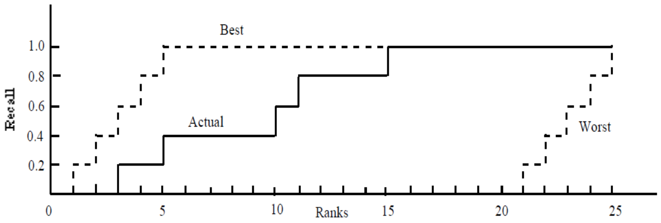

Figure 1. Illustration of how R_{norm} curve is bounded by the best and worst cases [15]

$$R_{norm} = 1 - \frac{\sum r_i - \sum i}{n(N-n)} \qquad \textbf{(2)}$$

where: ***r_i***: the rank at which the ith relevant document is retrieved, ***N***: collection size, and ***n***: number of relevant docs

Normalized recall can be seen as a good representative measure for recall-oriented IR applications. This measure is greater when all relevant documents are retrieved earlier. However it requires ranking the full collection. Applying R_{norm} on collections of huge numbers of documents is infeasible, since it is nearly impossible to rank a collection of potentially millions of documents. In addition, some relevant documents may have no match to a query leading to then not being retrieved at all. Calculating R_{norm} is impossible when some relevant documents are missed.

One approximation to address this problem is to consider any relevant documents not retrieved in the top N_{max} to be ranked at the end of the collection. Using this approximation to enable the calculation of R_{norm} leads to its value being nearly equal to the system recall at a cutoff of N_{max}. For example, for a collection of tens of thousands of documents and when retrieving the top 1000 documents; if recall @1000 equals 50%, R_{norm} with the previous approximation will equal 49.99% (Figure 2).

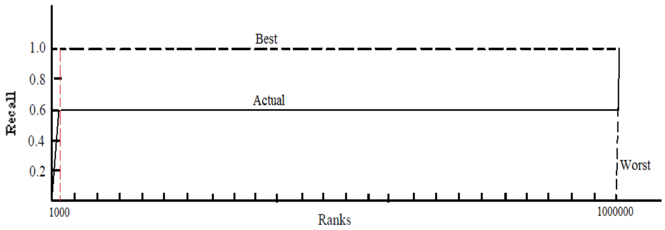

Figure 2. Illustration of how R_{norm} curve behaves with big collection of documents

5. PATENT RETRIEVAL EVALUATION SCORE (PRES)

In the previous sections we have demonstrated that current evaluation metrics do not represent system performance well in recall-oriented IR applications. In this section, a novel score is presented based on modifications to the normalized recall measure. As outlined in the previous section, R_{norm} can be seen as a good score for evaluating recall-oriented applications but only for small sized collection. Our new score "*Patent Retrieval Evaluation Score*" (PRES) is based on the same idea as the R_{norm} but with a different definition for the worst case. The new assumption for the worst case is to retrieve all the relevant documents just after the maximum number of documents to be checked by user (N_{max}). The idea behind this assumption is that getting any relevant document after N_{max} leads to it being missed by the user, and getting all relevant documents after N_{max} leads to zero recall, which is the theoretical worst case scenario. Applying this assumption in equation 2, N is replaced with $N_{max}+n$, where n is the number of relevant documents. Any relevant document non-retrieved in the top N_{max} is assumed to be the worst case (Figure 3). For example, for a retrieved ranked list for a topic with 10 relevant documents (n = 10) and for which the user is willing to check the top 100 documents (N_{max} = 100); the best case will be finding the 10 relevant documents in the ranks {1, 2, … 10}, and the worst case will be finding them in the ranks {101, 102, … 110}, which means the user missing all the relevant documents. Assuming retrieval of only 7 relevant documents in the top 100, then the missing 3 relevant documents will be assumed to be found at ranks {108, 109, 110}.

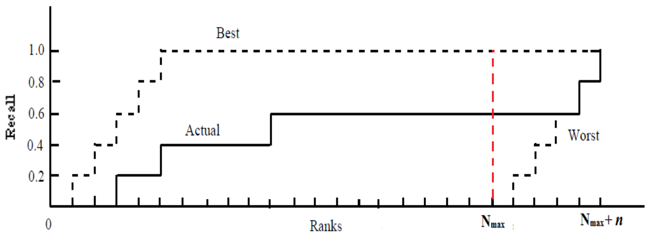

Figure 3. PRES curve is bounded between the best case and the new defined worst case

Equation 3 shows the calculation of PRES. Equation 4 shows the direct calculation of the summation of ranks of relevant documents in the general case, when some relevant documents are missing in the top N_{max} documents.

$$PRES = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{\max}} \qquad (3)$$

$$\sum r_i = \sum_{i=1}^{nR} r_i + nR\,(N_{\max} + n) - \frac{nR\,(nR-1)}{2} \qquad (4)$$

where, ***R***: Recall (number of relevant retrieved docs in the 1st N_{max} docs)

From equation 3, it can be inferred that PRES is a function of the recall of the system, the ranks of the retrieved documents, and the maximum number of results to be checked by user. For recall = R, the PRES value ranges from R, when retrieving all relevant document on the top of the list to nR^2/N_{max} when retrieving them at the bottom of the list.

6. ANALYSIS OF PRES PERFORMANCE

In this section, PRES is tested on the same sample examples as Table 1, with additional illustrative real samples from one run in the CLEF-IP 2009 patent retrieval task. In addition, the average performance is tested on real examples of 48 participants' runs from CLEF-IP 2009.

6.1. Performance with Sample Examples

Table 2. Performance of PRES with different IR systems

	Ranks of rel. docs	*AP*	*Recall*	PRES
System1	{1}	0.25	0.25	0.25
System2	{50, 51, 53, 54}	0.0481	1	0.51
System3	{1, 2, 3, 4}	1	1	1
System4	{1, 98, 99, 100}	0.2727	1	0.28

Table 2 shows how PRES performs with the sample examples presented in Table 1. From Table 2, it can be seen that PRES is a better representative measure for the system performance as a combination between system recall and average ranking of relevant documents. Some real samples of topics from one run of the CLEF-IP 2009 track are presented in Table 3 with maximum number of results to be checked by user N_{max} = 1000. In Tables 2 and 3, PRES is always less than or equal to recall, i.e. PRES is a portion of the recall depending on the quality of ranking the relevant documents relative to N_{max}. For example, getting a relevant document at the 10th rank will be very good when N_{max}=1000, good when N_{max}=100, but bad when N_{max} = 15, and very bad when N_{max}=10. Systems of higher recall can achieve a lower PRES value when compared to systems with lower recall but better average ranking. This is clear in Table 3, where the system with 67% recall achieves 63.6% PRES because of good ranking (41 and 54 among 1000), and the system with 100% recall achieves 52.5% for PRES because of the moderate ranking (60% of the relevant documents were found after the 500th rank among 1000).

Comparing PRES to *AP* for the samples in Table 3, it can be seen that AP is more sensitive to how early the first relevant document is found regardless of the number of documents to be checked by the user. However, PRES is more sensitive to the average ranking of the relevant retrieved documents as a whole relative to the maximum number of the documents the user is willing to check. The last sample topic in the table has a PRES of 96.43% even though it can be seen that the ranks are not in the top 10 or even 20 results. The reason is that N_{max}=1000, and the ranks {32, 35, 46} are considered relatively good to that number. Nevertheless, when calculating PRES with N_{max}=100, PRES value will be 64.33% which represents the average ranking of the relevant documents relative to the maximum number of documents to be checked.

Table 3. *AP*/*R*/PRES performance with real samples of topics

Ranks of rel. docs	*N*	*R*	AP	PRES
{98,296}	41	0.05	~ 0	0.039
{23,272,345}	6	0.5	0.01	0.394
{2,517,761}	6	0.5	0.085	0.288
{660,741}	3	0.667	0.001	0.201
{41,54}	3	0.667	0.021	0.636
{1,781}	3	0.667	0.334	0.407
{1,33,354,548,733,840,841}	7	1	0.157	0.525
{32,35,46}	3	1	0.051	0.964

6.2. PRES Average Performance

PRES was tested on 48 different submissions by 15 participants to the CLEF-IP 2009 Patent Track [16]. Table 4 shows the score for each submission in MAP, recall, and PRES. Participants IDs are anonymous and the number of topics for each participant used was 400 instead of the official 500 in order to further mask participant identities and to avoid violating the privacy of any of the participants. From the results, it can be seen that PRES reflects the recall with the average quality of the ranking, which is mainly reflected in the MAP. Run 21 (R21) which achieved the highest MAP and recall also achieved the highest PRES, and the same for the lowest ones. However, some submissions which achieved high precision but low recall were punished and received only a moderate PRES score. For systems which achieved high recall but low precision (which reflects bad ranking such as system R18), the PRES score was moderate too. Figure 5 plots the three scores od the same 48 submissions sorted by PRES from low to high. From Figure 5, it is noticed that PRES is a moderate score that can represent both the precision and recall of each run. Figure 6 shows the change in ranking of the submissions with the three scores. It can be seen that ranking using PRES is more biased to recall ranking, than MAP ranking. However, it is not always the case, for example R12 has moderate ranking in both recall and MAP, but lower ranking in PRES, which is due to the fact that MAP is more sensitive to the high ranking of some of the relevant documents, but PRES is dependent on relative average ranking of "All" relevant documents to N_{max}. From Figure 6, it can be seen that the scores have high agreement on the ranking of systems with very high or very low performances.

In order to check the agreement of the three scores, pair wise comparison of submissions was carried out with each two runs being compared: 1) 1st run is statistically significant better than 2nd run, 2) 2nd run is statistically significant better than 1st run, and 3) Both runs are statistically indistinguishable. Wilcoxon significance test with confidence level of 0.95 was used for comparing each of the two runs [6]. Comparing 48 runs in a pair wise manner led to 1,128 comparisons. The agreement of scores for each comparison is checked and plotted in Figure 7.

From Figure 6, it is clear that PRES is an intermediate score between recall and MAP. In addition, in a small number of cases (1%) PRES disagrees when recall and MAP agree. These situations are mainly for example when recall and MAP agree that system 1 (1st run) is better than system 2 (2nd run), but PRES shows that both systems have the same performance, or when recall and MAP agree that two systems are statistically indistinguishable, but PRES prefers one over the other.

Calculating the correlation between the ranking of the three scores, it is found that the scores are highly correlated in ranking, where the correlation between MAP and recall ranking is 0.71, PRES and recall ranking is 0.97, and PRES and MAP is 0.82; which means a 15% gain in correlation to MAP with very low loss in correlation to recall (3%). This shows the big advantage of PRES which is a recall-biased measurement with good reflection to the quality of ranking of relevant documents.

Table 4. MAP/Recall/PRES for 48 submissions in CLEF-IP

Run ID	MAP	Recall	PRES	Run ID	MAP	Recall	PRES
R01	0.077	0.530	0.434	R25	0.064	0.492	0.392
R02	0.087	0.617	0.499	R26	0.084	0.511	0.431
R03	0.084	0.609	0.497	R27	0.097	0.514	0.447
R04	0.053	0.219	0.213	R28	0.091	0.514	0.442
R05	0.000	0.020	0.011	R29	0.082	0.436	0.373
R06	0.000	0.016	0.009	R30	0.092	0.559	0.469
R07	0.000	0.012	0.007	R31	0.081	0.568	0.460
R08	0.000	0.016	0.009	R32	0.078	0.476	0.391
R09	0.071	0.454	0.369	R33	0.085	0.457	0.379
R10	0.088	0.533	0.430	R34	0.082	0.427	0.354
R11	0.087	0.489	0.404	R35	0.114	0.572	0.496
R12	0.088	0.534	0.430	R36	0.108	0.553	0.480
R13	0.065	0.508	0.406	R37	0.114	0.572	0.494
R14	0.068	0.467	0.363	R38	0.107	0.553	0.479
R15	0.064	0.434	0.348	R39	0.113	0.575	0.498
R16	0.020	0.197	0.148	R40	0.107	0.560	0.483
R17	0.067	0.584	0.463	R41	0.079	0.547	0.447
R18	0.033	0.656	0.490	R42	0.103	0.555	0.466
R19	0.105	0.600	0.529	R43	0.091	0.575	0.475
R20	0.003	0.051	0.040	R44	0.091	0.574	0.474
R21	0.266	0.760	0.691	R45	0.106	0.616	0.507
R22	0.028	0.256	0.200	R46	0.102	0.611	0.504
R23	0.087	0.728	0.603	R47	0.104	0.589	0.484
R24	0.011	0.069	0.054	R48	0.102	0.587	0.484

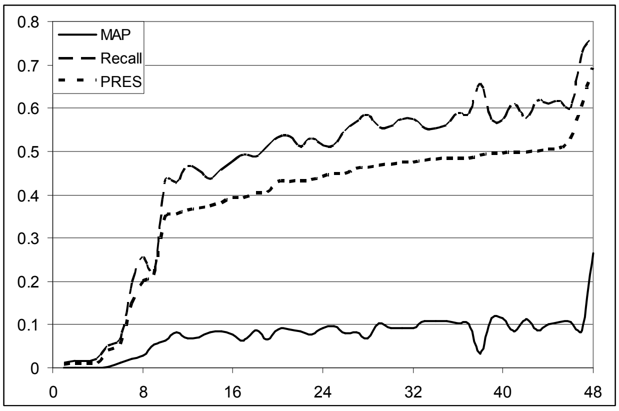

Figure 5. MAP/Recall/PRES for 48 submissions in CLEF-IP 2009 sorted by PRES

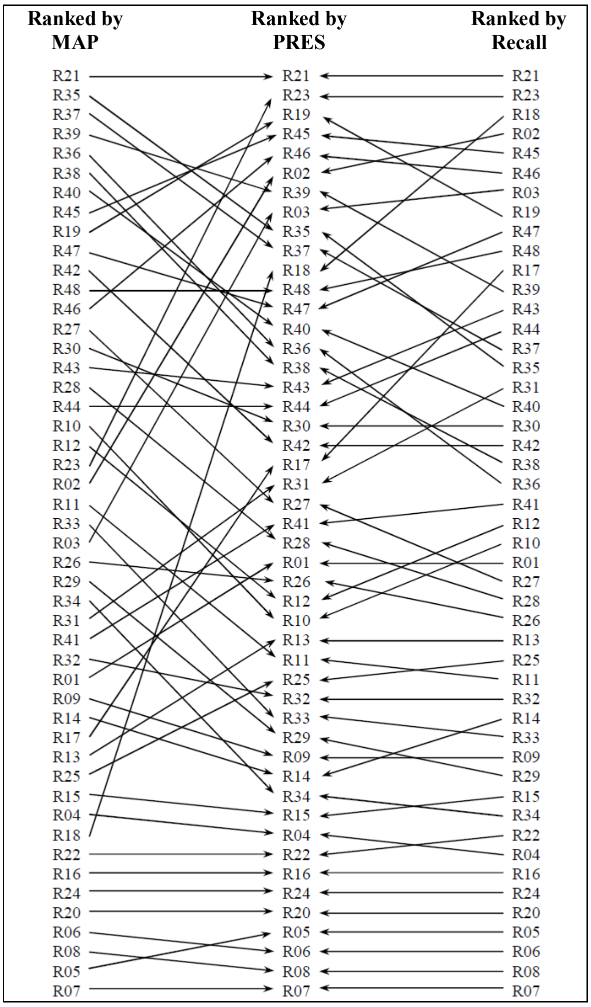

Figure 6. Ranking change of 48 submissions according to MAP/PRES/Recall

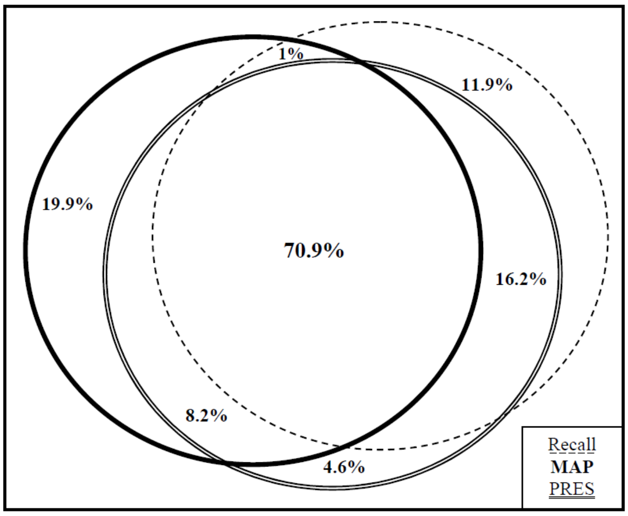

Figure 7. Agreement chart of MAP/Recall/PRES on pair wise comparison of 48 submissions

7. CONCLUSION & FUTURE WORK

In this paper, a study of patent retrieval evaluation has been described and a novel score "PRES" has been presented that is especially designed for this applications. The score has been tested and compared to the most widely used IR scores. Illustrative samples and real data examples demonstrated the effectiveness of the new score. The score reflects system recall combined with the quality of relative ranking of retrieved relevant documents within the maximum numbers of documents to be checked by the user. PRES value varies from R to nR^2/N_{max} according to the average quality of ranking of relevant documents; hence it can be seen as a function of system recall, ranking of relevant documents, and the maximum number of documents to be checked by a user (which directly affects the recall and relative ranking).

In future work, the utility of PRES as a measure for the patent retrieval could be investigated further by direct consultations with professional patent experts. Furthermore, the maximum number of documents to be checked by user (N_{max}) needs to be well identified based on realistic scenarios; The reason behind using N_{max}=1000 in the reported experiments is that it is the number used in the track, which does not mean it is the proper number to be used. Additionally, potential study for using MRR is suggested for topics that have relevant patents of type X, this type of relevant patents totally invalidate the novelty of patent application and hence, one it is found, the examiner doesn't have to continue search for relevant patents. However, this type of data is not available for us right now.

8. ACKNOWLEDGMENT

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project

9. REFERENCES

1. Ali M S., M. P. Consens, G. Kazai, M. Lalmas, Structural relevance: A common basis for the evaluation of structured document retrieval, *in CIKM '08: Proceeding of the 17th ACM CIKM.* (2008)
2. Azzopardi L., M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR '07,* pages 455–462, New York, NY, USA, 2007. ACM. (2007)
3. Baeza-Yates J. and B. Ribeiro-Neto. Modern Information Retrieval. (1999)
4. Buckley C., D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling. In *SIGIR '06,* pages 619–620, New York, NY, USA, 2006. ACM. (2006)
5. Carterette B., P. N. Bennett, D M. Chickering and S. T. Dumais. Here or There: Preference Judgments for Relevance. *ECIR 2008*, 2008.
6. D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *SIGIR '93, pages 329–338, New York, NY, USA*, (1993)
7. Fujii, A., Iwayama, M., and Kando, N. Overview of patent retrieval task at NTCIR-4. In *Proceedings of the fourth NTCIR workshop on evaluation of information retrieval, automatic text summarization and question answering, June 2–4, Tokyo, Japan*. (2004)
8. Fujii A., M. Iwayama, and N. Kando. Overview of the patent retrieval task at the NTCIR-6 workshop. *In Proceedings of*

the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pages 359–365. (2007)

9. Graf E. and L. Azzopardi, A methodology for building a patent test collection for prior art search, Proceedings of the Second InternationalWorkshop on Evaluating Information Access (EVIA). (2008)
10. Iwayama M., A. Fujii, N. Kando, and A. Takano. Overview of patent retrieval task at NTCIR-3. *In Proceedings of the Third NTCIR Workshop on evaluation of information retrieval, automatic text summarization and question answering*. (2003)
11. Jordan C., C. Watters, and Q. Gao. Using controlled query generation to evaluate blind relevance feedback algorithms. In *JCDL '06* pages 286–295, New York, NY, USA, (2006)
12. Kamps J., Jovan Pehcevski , Gabriella Kazai , Mounia Lalmas , Stephen Robertson. INEX 2007 evaluation measures. In *INEX 2007 Dagstuhl Castle, Germany*. (2007)
13. Leong M.K. Patent Data for IR Research and Evaluation. *In Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pages 359–365*. (2001)
14. Oard, D.W., Hedin, B., Tomlinson, S., Baron, J.R.: Overview of the TREC 2008 legal track. In: *Proceedings TREC 2008*. (2009)
15. Rijsbergen, C.J. *v.: Information Retrieval, 2nd edition. Butterworth- Heineman.* (1979)
16. Roda G., Tait J., Piroi F., and Zenz V. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. *CLEF working notes* 2009, Corfu, Greece, (2009)
17. Tague J., M. Nelson, and H. Wu. Problems in the simulation of bibliographic retrieval systems. In *SIGIR '80: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 236–255, Kent, UK, UK, 1981. Butterworth & Co. (1981)
18. Voorhees E. M. and D. M. Tice. The TREC-8 Question Answering Track Evaluation. In *Text Retrieval Conference TREC-8*, 1999. (1999)
19. Voorhees, E. M. The TREC robust retrieval track. In: *ACM SIGIR Forum* 39 (1) pp. 11-20. (2005)
20. Xue X. and Croft W. B. Automatic Query Generation for Patent Search. In *Proceeding of CIKM'09, November 2–6, 2009, Hong Kong, China*. (2009)
21. Zhu J. and Tait J. A proposal for chemical information retrieval evaluation. *Proceeding of the 1st ACM workshop on Patent information retrieval*. CIKM 2008. (2008)