Examining the Robustness of Evaluation Metrics for Patent Retrieval with Incomplete Relevance Judgements

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. Recent years have seen a growing interest in research into patent retrieval. One of the key issues in conducting information retrieval (IR) research is meaningful evaluation of the effectiveness of the retrieval techniques applied to task under investigation. Unlike many existing well explored IR tasks where the focus is on achieving high retrieval precision, patent retrieval is to a significant degree a recall focused task. The standard evaluation metric used for patent retrieval evaluation tasks is currently mean average precision (MAP). However this does not reflect system recall well. Meanwhile, the alternative of using the standard recall measure does not reflect user search effort, which is a significant factor in practical patent search environments. In recent work we introduce a novel evaluation metric for patent retrieval evaluation (PRES) [13]. This is designed to reflect both system recall and user effort. Analysis of PRES demonstrated its greater effectiveness in evaluating recall-oriented applications than standard MAP and Recall. One dimension of the evaluation of patent retrieval which has not previously been studied is the effect on reliability of the evaluation metrics when relevance judgements are incomplete. We provide a study comparing the behaviour of PRES against the standard MAP and Recall metrics for varying incomplete judgements in patent retrieval. Experiments carried out using runs from the CLEF-IP 2009 datasets show that PRES and Recall are more robust than MAP for incomplete relevance sets for this task with a small preference to PRES as the most robust evaluation metric for patent retrieval with respect to the completeness of the relevance set.

1 Introduction

Interest in patent retrieval research has shown considerable growth in 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 of these tasks at these workshops is to encourage researchers to explore patent retrieval on common tasks in order understand the issues in providing effective patent retrieval and to establish the best IR methods for doing this. Due to the important of finding all items relevant to the query, patent retrieval is generally identified as a recall-oriented retrieval task [7]. While in practice achieving 100% recall may not be achievable, the operational objective of patent retrieval is to maximise search recall for the minimum effort by the user, typically a professional

patent examiner. In contrast to this, the majority of IR evaluation tasks are precision focused IR tasks, where one or two of the relevant documents are often sufficient to achieve user satisfaction by addressing their information need. In this latter case the objective is to find these relevant documents as quickly as possible at high rank in a retrieved document list to minimise user effort. While considerable effort has been devoted to the study of evaluation metrics for precision focused IR tasks, the evaluation of patent retrieval is still an open area of research due to the special objective of the search task. The standard evaluation metric used for most IR tasks remains mean average precision (MAP). While patent retrieval is recognised to be a recall-oriented IR task, MAP is still the most widely used score for patent retrieval evaluation. In previous work we conducted a careful examination of the suitability of MAP for evaluating patent retrieval [13]. As a result of this investigation we introduced a new patent retrieval evaluation score (PRES) as a new evaluation metric specifically designed for recall-oriented IR tasks.

Laboratory evaluation for IR tasks such as those examined at TREC, CLEF and NTCIR relies on a standard model of a target document collection, representative search topics and a relevance set indicating which documents from the document set are relevant to each topic. An important practical issue is that the size of realistic document collections means not all documents can be assessed for relevance to each topic. An approximation to the true complete relevance is made based on some reasonable method for example using a pooling procedure involving manual relevance assessment of a selected subset of the document set or by assuming a relevance relationship between the topic and a subset of the documents [6]. In consequence the set of known relevant items for each topic is almost certainly incomplete. An important question which has been examined in a number of studies is the extent to which experimental results for IR tasks are robust to the design of the test collections [3, 4, 17]. One aspect of this research has focused on the number of topics to be used for an IR experiment to achieve reliable and robust results [4, 17], while others have focused on how evaluation metrics are stable with different values for cut-off levels of the retrieved results list [4]. Other work has studied the robustness of evaluation scores relating to the incompleteness of relevance judgements [2]. This latter study examined the stability of ranking system effectiveness when using different portions of the relevance judgements. Some score metrics have been developed to overcome problems relating to incomplete relevance assessment, such as inferred average precision (infAP) [1] and Bpref [5]. Although these metrics have proven some robustness to incomplete judgments, they are focused on the precision of retrieval results, which is not the objective of the recall-oriented patent retrieval task. An important and previously unexplored issue is the examination of the behaviour of evaluation metrics with incomplete judgements studies when evaluating a recall-oriented IR application. Similarly to standard precision-orientated tasks, such as ad hoc search, relevance sets in patent retrieval will generally be incomplete. Thus an important question is the stability of evaluation metrics across different IR systems used for patent search variations in the incomplete relevance set. In the case of patent retrieval, the practical significance of this is to assess the extent to which the value of evaluation metric measured using an incomplete relevance set is a prediction of its ability to help the patent examiner to find further relevant documents if they continued to search in the collection.

This paper provides what we believe to be the first study of the robustness of evaluation metrics for patent retrieval with incomplete relevance judgements. We compare the behaviour of standard MAP and Recall and with our PRES metric designed for the patent retrieval task. We compare behaviour of these three metrics since the first two are the ones most commonly used to evaluate effectiveness of a patent retrieval system, and the third is specifically designed for evaluation of patent retrieval and has shown desirable behaviour in terms of evaluating submissions to the CLEF-IP task [13]. Our investigation uses similar techniques to those described in [3] to test the robustness of these three metrics, although some modification is required due to the small number of relevant documents per topic. Experiments were performed on 48 runs submitted by participants in the CLEF-IP 2009 task [16]. Results show strong robustness for PRES and Recall for this type of task, while MAP is shown to have much weaker robustness for MAP to variations in the known relevance set. Our conclusion is that these results indicate that MAP is not a suitable evaluation metric for recall-oriented IR tasks in general and patent retrieval in particular, especially in the absence of a guarantee of the completeness of the relevance judgements.

The remainder of this paper is organized as follows: Section 2 provides background on patent retrieval and the evaluation metrics used for this task, Section 3 describes our experimental setup and provides relevant details of the CLEF-IP 2009 task, Section 4 describes the results of our investigation of the robustness of the evaluation scores, and finally Section 5 concludes and provides possible direction for future research.

2 Background

This section provides an introduction to the task of patent retrieval and reviews the history of its introduction in the NTCIR and CLEF IR evaluation campaigns. In addition, the evaluation metrics commonly used to evaluate patent retrieval are described in summary.

2.1 Patent Retrieval

Evaluation of patent retrieval was proposed in NTCIR-2 in 2001 [12]. Since then patent retrieval has featured as a track in all NTCIR[1] campaigns. Patent retrieval was introduced much more recently at CLEF[2] in 2009, as the CLEF-IP (CLEF Intellectual Property) track [16]. Patent retrieval is of interest in IR research since it is of commercial interest and is a challenging IR task with different characteristics to popular IR tasks such as precision-orientated ad hoc search on news archives or web document collections [12, 16]. Various tasks have been created around patents; some are IR activities while others focus on tasks such as data mining from patents and classification of patents.

[1] http://www.nii.ac.jp/
[2] http://www.clef-campaign.org/

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

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].

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.

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].

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 parts in common with the current patent [16]. In this type of task, patent citations are considered as the relevant documents, and the objective is to automatically find these patent citations. These citations are usually identified by a patent office and take considerable periods of time to search for them manually [9, 16].

In this paper, experiments examine prior-art search task for patent retrieval. This kind of task is characterized by the small number of relevant documents per topic. This small number of relevant items is to be expected since any filed patent should contain novel ideas that typically should not be contained in prior patents. Although the fact that these citations take a huge amount of effort and time to identify, there is no guarantee that the recall of finding all the relevant documents to be 100%[3]. Hence, the focus of the study in this paper is to examine the impact on evaluation metrics of missing any of these relevant documents.

2.2 Evaluation Metrics for Patent Retrieval

Mean Average Precision (MAP). While many evaluation metrics have been proposed for ad hoc type IR tasks, by far the most popular in general use is MAP [1]. The standard scenario for use of MAP in IR evaluation is to assume the presence of a collection of documents representative of a search task and a set of test search topics (user queries) for the task along with associated manual relevance data for each topic. For practical reasons the relevance data is necessarily not exhaustive, but the relevance data for each topic is assumed to be a sufficient proportion of the relevant documents from the document collection for this topic. "Sufficient" relates to the reliability of the relative ranking of the IR systems under examination. Several techniques are available for determining sufficient relevant documents for each topic [6]. As its name implies, MAP is a precision metric, which emphasizes returning

[3] Based on discussions with staff from the European Patent Office.

relevant documents earlier in a ranked list (Equation 1). The impact on MAP of locating relevant documents later in the 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 give a meaningful interpretation for recall-focused tasks. Despite this observation, MAP remains the standard evaluation metric used for patent retrieval tasks.

$$Avg.\ precision = \frac{\sum_{r=1}^{N}(P(r) \times rel(r))}{n} \quad (1)$$

where ***r*** is the rank, *N* the number of documents retrieved, *rel(r)* a binary function of the document relevance at a given rank, *P(r)* is precision at a given cut-off rank *r*, and *n* is the total number of relevant documents (|{*relevant documents*}|).

Recall. Recall is a standard score metric measuring the ability of an IR system to retrieve relevant documents from within a test collection. It gives no indication of how retrieved relevant documents are ranked within the retrieved ranked list. For this reason recall is generally used in conjunction other evaluation metrics, typically precision, to give a broader perspective on the behaviour of an IR system. Although recall is the main objective in patent retrieval, it has not been used as a performance score to rank different systems due to its failure to reflect the quality of ranking of the retrieved results. Nevertheless, recall remains a very important metric to show this aspect of a patent retrieval system's behaviour which is not reflected by precision measures such as MAP.

Patent Retrieval Evaluation Score (PRES). Based on a review of the objectives of patent retrieval and the deficiencies existing metrics, in earlier work we introduced PRES as a novel metric for evaluating recall-oriented IR applications [13]. PRES is derived from the normalized recall measure (R_{norm}) [15]. It measures the ability of a system to retrieve all known relevant documents earlier in the ranked list. Unlike MAP and Recall, PRES is dependent on the relative effort exerted by users to find relevant documents. This is mapped by N_{max} (Equation 2), which is an adjustable parameter that can be set by users and indicates the maximum number of documents they are willing to check in the ranked list. PRES measures the effectiveness of ranking documents relative to the best and worst ranking cases, where the best ranking case is retrieving all relevant documents at the top of the list, and the worst is to retrieve all the relevant documents just after the maximum number of documents to be checked by the 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. Figure 1 shows an illustrative graph of how to calculate PRES, where PRES is the area between the actual and worst cases (A_2) divided by the area between the best and worst cases (A_1+A_2).

N_{max} introduces a new definition to the quality of ranking of relevant results, as the ranks of results are relative to the value of N_{max}. For example, getting a relevant document at rank 10 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 with higher Recall can achieve a lower PRES value when compared to systems with lower Recall but better average ranking. The PRES value varies from R to nR^2/N_{max}, where R is the Recall, according to the average quality of ranking of relevant documents.

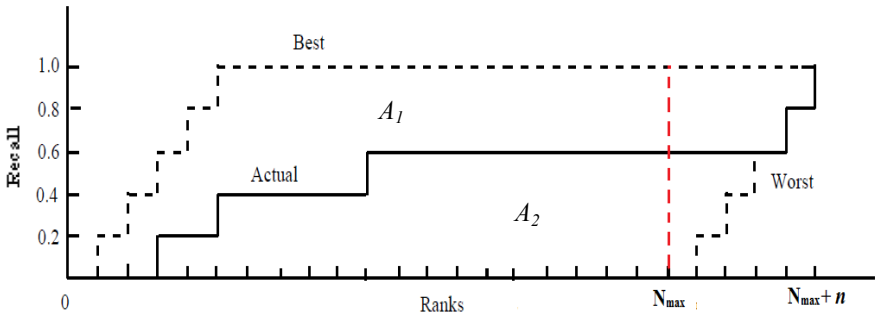

Fig. 1. PRES curve is bounded between the best case and the new defined worst case

$$PRES = \frac{A_2}{A_1 + A_2} = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{\max}} \quad (2)$$

where r_i is the rank at which the ith relevant document is retrieved, N_{max} is the maximum number of retrieved documents to be checked by the user, i.e. the cut-off number of retrieved documents, and n is the total number of relevant documents.

In [13] we demonstrate that PRES is a more suitable score metric for ranked recall-oriented IR applications than existing evaluation metrics. This was illustrated using participants' result runs submitted to the CLEF-IP 2009 which have been released for research purposes. This earlier study did not explore the robustness of the compared evaluation measures with respect to the completeness of the relevance set. The study in this paper extends our existing work to this important previously overlooked dimension of the behaviour of the metrics.

In the remainder of this paper describes our investigation is into the stability of MAP, Recall, and PRES for the same experimental data in [13] using a similar experimental approach to that used in [3].

3 Experimental Setup

The study in this paper is performed on the CLEF-IP 2009 patent retrieval task [16]. The submitted runs for the main task in CLEF-IP 2009 are used to compare the robustness of evaluation metrics when relevance judgements are incomplete.

3.1 CLEF-IP 2009 Track

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. The collection to be searched contains about one million patents filed in the period from 1985 to 2000 [16]. The main task in the track was prior-art search (section 2.1); where the objective is to automatically retrieve all cited patents found in the collection. These citations are originally identified by the patent applicant or the patent office.

Forty-eight runs were submitted by the track participants. The track organizers have been kind enough to release these runs in order to encourage investigation of new evaluation methodologies for patent retrieval. Each run consists of up to the top 1000 ranked results for each topic. The topic set consists of 500 topics, which is the smallest topic set provided by the track. Different topic sets of sizes up to 10,000 topics were available. However the number of runs submitted for these topic sets was much less than 48. In our previous research developing PRES and comparing it to MAP and Recall with these 48 runs, only 400 topics were used in order to anonomize IDs of the runs of the individual participants [13]. The same set of runs using the 400 topic subset of each run is used in this investigation. The average number of relevant documents per topic is 6 [16] with a minimum number of 3. This is considerably less than is typically found for existing standard ad hoc IR evaluation tasks. Figure 2 shows the number of relevant documents per each topic for the 400 topic set used in the experiments. Figure 2 shows that more than 25% of the topics have only 3 relevant documents. This small number emphasizes the importance of the robustness of the evaluation metric used since missing even one relevant document can have a huge impact on the relationship between the topic and the collection.

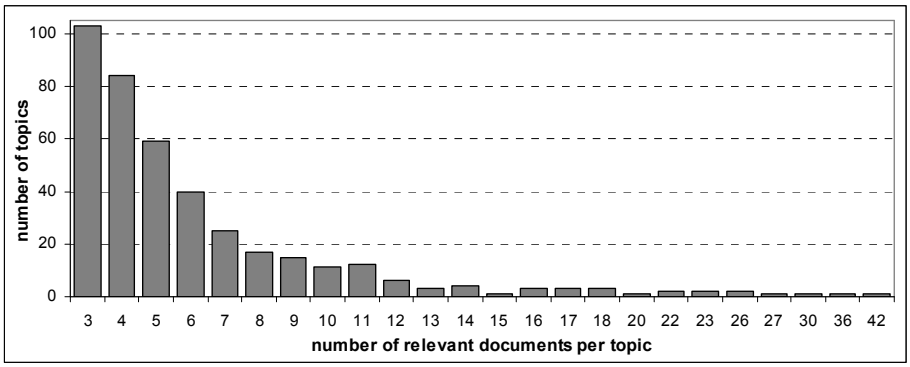

Fig. 2. Distribution of the number of relevant documents per topic in the test collection (average number of relevant documents = 6)

3.2 Experimental Setup

In order to check the robustness of the evaluation metrics, several versions of the relevance judgements *qrels* were created by generating different fractions of the judgements *f-qrels*. The original *qrels* provided by the track are assumed to be the full

100% *qrels*[4]; other versions representing fractions of the *qrels* were generated by selecting a certain fraction value (*f*) of the relevant documents for each topic in the topic set. In [3], a similar setup was used to the study in this paper, but with two main differences. The first is that this study is focusing on a recall-oriented patent retrieval task, where missing any relevant document in the assessments should be considered very harmful for a fair evaluation of systems. The second difference lies in the nature of the data collections used in the studies. In [3] TREC collections characterized by a high average number of relevant documents per topic were used. This allowed the study to test many *f-qrels*, where *f* ranged from 0.01 to 0.9. In the current experiments, patent collection is used, which is characterized by the relatively low average number of relevant documents per topic. This low number does not allow such large variation in the values of *f*.

To conduct the robustness experiments, four fraction values of the *qrels* were used (*f* = 0.2, 0.4, 0.6, and 0.8). For each *f* value, three *f-qrels* are generated, where the selection of the fraction of relevant documents was randomized, hence the three versions are always different. This produced a set of 12 *f-qrels*. The objective is to compare the ranking of the runs according to each score using these *f-qrels* to the ranking when using the full *qrels*. Kendall's tau correlation [11] was used to measure the change in the ranking. The higher the correlation for smaller values of *f*, the more robust the metric is to the incompleteness of relevance judgements.

Two values of cut-offs were used, the first is the one reported in the CLEF-IP track itself which is 1000 results for each topic. The second cut-off value is 100, which is more realistic for a patent retrieval task since this is the order of the number of documents typically checked for relevance by a patent examiner for each topic.

4 Results

Table 1 shows the Kendall tau correlation values for the three scores at different cut-offs and for the different samples for each value of *f* (% *qrels*). Figures 3 and 4 plot the worst-case values of the correlation for the three scores for cut-offs values of 100 and 1000 respectively. Table 1 and Figures 3 and 4 demonstrate several points:

- MAP has a much lower Kendall tau correlation when compared to the Recall and PRES, especially for lower values of *f*. This result surprisingly shows that the most commonly used metric for patent retrieval evaluation is the least reliable one when there is no guarantee of the completeness of the relevance judgements.
- Recall and PRES have nearly symmetric performance with incomplete judgements with slightly better performance to PRES for lower values of cut-off.
- Following the study of Voorhees [18] which determines rankings to be nearly equivalent if they have a Kendall tau correlation value of 0.9 or more, and to have a noticeable difference for Kendall tau correlation less than 0.8.

[4] Although of course this is actually known not be the case since exhaustive manual analysis of the collection has not been carried out.

According to the results found in our investigation, Recall and PRES will have an equivalent ranking for systems even with only 20% of the relevance judgements. However, MAP may have a noticeable change of system ranking even if only 20% of the judgements are missing.

- A drop in the curve of correlation of MAP for cut-off of 100 can be seen from Figure 3 when % *qrels* = 60%. One explanation for this can be the randomness used in selecting the fraction of relevant document from the *qrels*.

Although PRES and Recall have similar performance with the incomplete judgements, both metrics cannot be claimed to be the same. The experiments in this paper only test one aspect of the evaluation metrics. However, additional factors should be taken into consideration when considering the suitability of an evaluation metric, in this case the metric's ability to distinguish between the performances of different systems in a fair way. Bearing all factors in mind, PRES can be considered as the more suitable evaluation metric for patent retrieval since it has been shown to have a greater ability to rank systems in a recall-oriented IR environment [13]. Both this and the findings in this paper with regard to the consistent performance of PRES across different fractions of the relevance judgements recommend the use of PRES for this type of IR application.

Table 1. Correlation between the ranking of 400 topics from 48 runs with different percentages of incomplete judgements and different cut-offs for MAP, Recall, and PRES

		Cut-off = 100			**Cut-off = 1000**		
% *qrels*	**Sample**	**MAP**	**Recall**	**PRES**	**MAP**	**Recall**	**PRES**
20%	1	0.50	0.94	0.94	0.58	0.93	0.93
	2	0.71	0.88	0.92	0.70	0.92	0.89
	3	0.59	0.90	0.90	0.66	0.90	0.90
	avg.	0.60	0.90	0.92	0.65	0.92	0.91
40%	1	0.93	0.92	0.93	0.76	0.96	0.96
	2	0.71	0.93	0.93	0.87	0.95	0.93
	3	0.75	0.91	0.92	0.84	0.94	0.92
	avg.	0.79	0.92	0.93	0.82	0.95	0.94
60%	1	0.66	0.95	0.96	0.82	0.98	0.98
	2	0.66	0.95	0.97	0.82	0.98	0.97
	3	0.65	0.96	0.96	0.90	0.97	0.98
	avg.	0.66	0.95	0.97	0.85	0.98	0.97
80%	1	0.79	0.96	0.97	0.96	0.98	0.98
	2	0.94	0.97	0.99	0.94	0.97	0.98
	3	0.75	0.98	0.98	0.84	0.98	0.99
	avg.	0.83	0.97	0.98	0.91	0.98	0.98
100%	NA	1	1	1	1	1	1

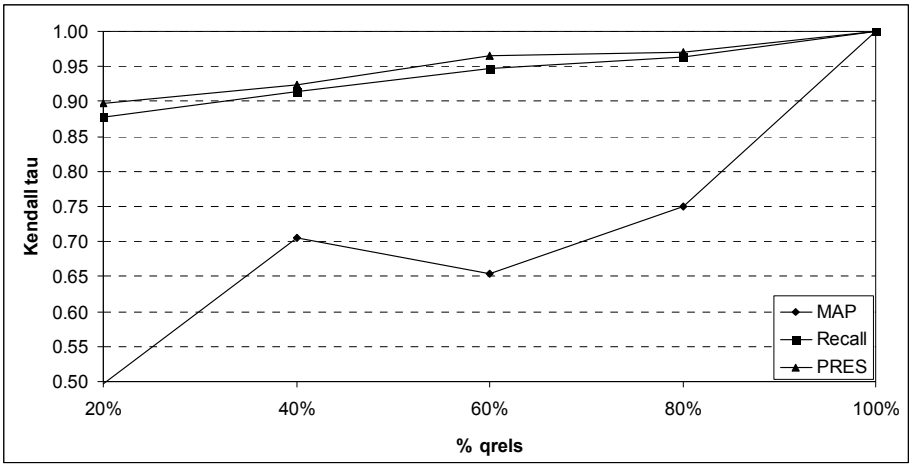

Fig. 3. Lowest Kendall tau correlation values for MAP/Recall/PRES for cut-off = 100

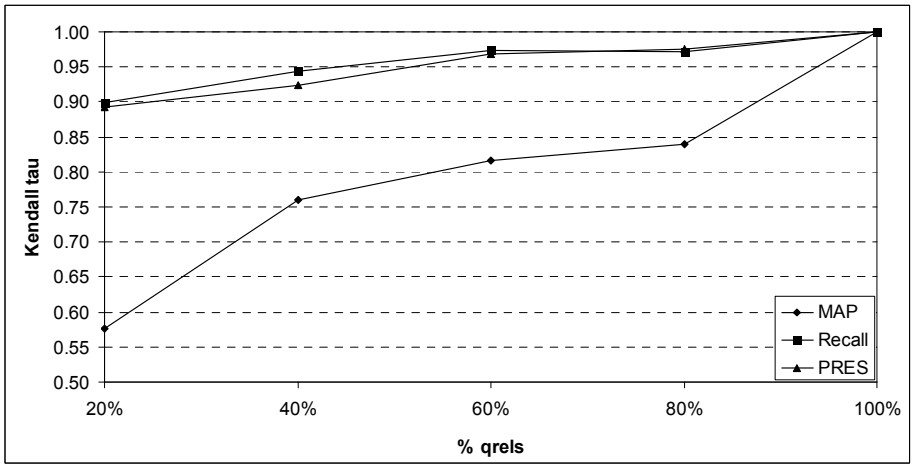

Fig. 4. Lowest Kendall tau correlation values for MAP/Recall/PRES for cut-off = 1000

5 Conclusion and Future Work

In this paper, a study for the robustness of the evaluation metrics used for the recall-oriented patent retrieval task has been presented. The aim of the study was to test the consistency of the performance of three evaluation metrics which are currently used for patent retrieval evaluation (MAP, Recall, and PRES) when the relevance judgement set is incomplete. Different fractional values of the *qrels* with different samples were used to conduct the experiments. Kendall tau correlation was used as the measure of the consistency of the ranking of systems. Results show that the most

commonly used score for evaluating patent retrieval, MAP, is the least reliable evaluation metric to be used in this kind of IR application, since it shows the least consistency in ranking different runs when the relevance judgements are incomplete. PRES and Recall both have very robust performance even when only small portions of the relevant judgements are available. Considering the strong performance of PRES for evaluating recall-oriented IR applications, in addition to the results of this paper; PRES can be recommended as a standard score metric for evaluating recall-oriented IR application, especially patent retrieval.

As future work, this study can be extended to include more runs submitted to the upcoming CLEF-IP tracks. PRES is due to be considered as one of the metrics used for evaluating this task in CLEF-IP in 2010. It will be interesting to see whether these new runs provide further evidence for the robustness of these metrics. Furthermore, the study could be expanded to capture other types of evaluation metrics, such as geometrical mean average precision (GMAP), precision at different cut-off values, and normalized discount cumulative gain (NDCG). Although this kind of study can be interesting from the robustness point of view, there is always another dimension which needs to be considered when selecting an evaluation metric namely, what feature of a system‘s behaviour is the metric evaluating.

6 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 at Dublin City University.

7 References

1. Aslam J. A., and E. Yilmaz. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management CIKM,* Arlington, Virginia, USA, page102-111, (2006)
2. Baeza-Yates, J., and B. Ribeiro-Neto. Modern Information Retrieval, *Addison Wesley*. (1999)
3. Bompad, T., Chang, C.-C., Chen, J., Kumar, R., and Shenoy, R.: On the robustness of relevance measures with incomplete judgements. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, Amsterdam, The Netherlands, ACM. (2007)
4. Buckley, C., and Voorhees, E.: Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athens, Greece, pages 33-40, 2000. ACM. (2000)
5. Buckley, C., and Voorhees, E. M. Retrieval evaluation with incomplete information. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, South Yorkshire, UK, pages 25-32, (2004)

6. Buckley, C., Dimmick, D., Soboroff, I., and Voorhees, E.: Bias and the limits of pooling. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, WA, USA, pages 619–620, ACM. (2006)
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*, Tokyo, Japan. (2004)
8. Fujii, A., Iwayama, M., and Kando, N.: 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*, Tokyo, Japan, pages 359–365. (2007)
9. Graf, E., and Azzopardi, L.: A methodology for building a patent test collection for prior art search. In Proceedings of The Second International Workshop on Evaluating Information Access (EVIA 2008), Tokyo, Japan. (2008)
10. Iwayama M., Fujii, A., Kando, N., and Takano, A.: Overview of patent retrieval task at NTCIR-3. In *Proceedings of the 3rd NTCIR Workshop on evaluation of information retrieval, automatic text summarization and question answering*. Tokyo, Japan (2003)
11. Kendall, M.: A new measure of rank correlation. *Biometrika*, 30(1/2):81-93, 1938.
12. 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*, Tokyo, Japan, pages 359–365. (2001)
13. Magdy, W., and Jones, G. J. F.: PRES: a score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the 33rd annual international ACM SIGIR conference on Research and development in information retrieval*. Geneva, Switzerland, ACM. (2010)
14. Van Rijsbergen, C. J. Information Retrieval, 2nd edition. *Butterworths*. (1979)
15. Robertson S. E. The parametric description of the retrieval tests. *Part 2: Overall measures. Journal of Documentation*, 25(2):93-107, (1969)
16. Roda, G., Tait, J., Piroi, F., and Zenz, V.: CLEF-IP 2009: Retrieval experiments in the Intellectual Property domain. *In CLEF 2009 Working Notes,* Corfu, Greece. (2009)
17. Voorhees E. M.: Special Issue: The Sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1). (2000)
18. Voorhees E. M.: Evaluation by highly relevant documents. In *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, U.S.A., pages 74–82. (2001)