Tolerance to Irrelevance: A User-effort Oriented Evaluation of Retrieval Systems without Predefined Retrieval Unit

Arjen P. de Vries
CWI
Amsterdam
The Netherlands

Gabriella Kazai
Queen Mary University London
London
United Kingdom

Mounia Lalmas
Queen Mary University London
London
United Kingdom

Abstract

Video and XML retrieval test collections call for evaluation metrics that do not require a predefined retrieval unit. The use of traditional recall and precision metrics is problematic due to issues caused by 'overlap' between result and reference items. The paper proposes evaluation metrics derived from a user-effort oriented view of information retrieval to address these problems. It builds on the Expected Search Length metric of Cooper, revived by Dunlop for the Expected Search Duration metric. Our work extends these previous works by demonstrating how to handle systematically the overlap problems introduced when the assumption of a fixed, predefined retrieval unit is removed from the benchmark setting.

Keywords: structured document retrieval, video retrieval, evaluation metrics.

1. Introduction

Evaluation experiments in information retrieval (IR) are traditionally based on standardised test collections, which typically provide a fixed set of documents, user requests and relevance assessments allowing to best focus on the retrieval approaches to compare their relative effectiveness (Tague-Sutcliffe, 1992). This so-called *Cranfield tradition* of experimental evaluation has converged in over twenty years into what is now known as 'standard IR evaluation practise'. It has become universal through the retrieval evaluations organised at the Text REtrieval Conference (TREC) (Harman, 1992).

Since the start of TREC in 1992, several new media types and retrieval problems have emerged. The evaluation of systems that aim to tackle these retrieval tasks has become the focus of new evaluation tracks, widening the scope of TREC. One such track deals with the problem of video retrieval, aiming to promote progress in content-based retrieval from digital video. Participating groups index a publicly available video collection and return ranked lists of *video clips* for a set of topics. The boundaries for the units of video to be retrieved are not predefined and each system makes its own independent judgement of what fragment of a video programme constitutes a relevant result item.

Another domain that has received increased interest recently is the retrieval of XML documents. Exploiting the explicitly available knowledge of document structure, XML retrieval systems aim at retrieving document components instead of whole documents. Their task is similar to that of video retrieval systems in the sense that they both focus the user's attention to relevant fragments within the traditional units of retrieval. The evaluation of XML retrieval systems has been set as the goal of the INitiative for the Evaluation of XML retrieval (INEX) (Kazai, Lalmas, Fuhr, & Gövert, 2004).

Given a test collection, the *de facto* standard for quantifying search system performance is to evaluate a system's effectiveness by using the combination of recall and precision. Measuring the set-based recall and precision, however, requires a *predefined* unit of retrieval, based on which, the elements that make up the retrieved and relevant sets are defined. The assumption that the entity to be retrieved can be defined a priori is however violated in the 'new' video and XML retrieval problems. From the user's point of

view, a ranked list of lengthy articles or continuous media results is not easily scanned for relevance. Inspecting a result set for relevance is a time-consuming task, the whole of the article must be read or the video must be viewed. The evaluation methodology should take this into account by measuring at varying levels of granularity instead of full documents.

In the case of video retrieval, the granularity is typically based on 'shots' or 'scenes'. Given a digital video, which is organised in frames (usually 25 or 30 per second), a shot is defined as a sequence of frames recorded contiguously, usually ended with a camera cut or an edit special effect. A scene is a group of consecutive shots that shares semantics in terms of time, place, objects or events. Because only the shot boundaries can be detected automatically without making too many mistakes, most video retrieval systems segment video data at the shot level.

In XML retrieval, the nested structure of the XML tags allows for varying levels of granularity, where the selection of the 'right' granularity to present to the user is represented by the most specific, relevant document components. In addition to the granularity levels provided by the XML markup, the boundaries of sentences (or even words) can serve to mark a smallest unit of retrieval.

In both cases, the retrieval systems decide the granularity of the target entities to be presented to the user, a property of what we call the 'retrieval unit'. Given that the standard approach to evaluation, using the combination of recall and precision, depends on a predefined retrieval unit, the evaluation of a video or XML retrieval system's effectiveness poses new challenges. This paper addresses the question of how to evaluate systems using test collections when the assumption of a known, predefined retrieval unit no longer holds.

We base our investigations on a user-effort oriented view of information retrieval and explore the use of alternative measures to recall and precision. In particular, we make use of Cooper's Expected Search Length (ESL) metric, which provides a measure of user effort by predicting the expected number of documents the user must read before finding a desired number of relevant documents (Cooper, 1968). Dunlop suggested to extend this measure to predict the time the user needed to process the documents, and called this Expected Search Duration (ESD) (Dunlop, 1997). The primary contribution of our work is to extend the work by Cooper and Dunlop to address the overlap problems introduced when the systems evaluated do not assume a fixed, predefined retrieval unit.

The paper is organised as follows. We first summarise the main properties of the TREC-10 video collection, and the INEX'03 XML document collection. Section 3 discusses the effect that allowing system-determined retrieval units has on IR evaluation. Section 4 argues how the resulting varying retrieval unit size leads naturally to evaluation metrics that model user effort. It explains how to abstract from the user, allowing retrieval system evaluation using laboratory tests without predefining the retrieval unit size. Section 5 explains how to instantiate this generic abstract approach to obtain evaluation metrics for both video and XML retrieval. The paper concludes by identifying some limitations to be removed in further research.

2. Test Collections

TREC-10 introduced an experimental Video Track, now usually referred to as TRECVID 2001, with two search tasks: known-item(s) searches reflecting a specialised type of user need and general searches expressing general statements of information need. The collection consists of approximately twelve hours of MPEG-1 encoded video (totalling over 6 Gb), from 85 video programmes (videos) usually of documentary nature but varying in age, production style, and quality. The topics are truly multimedia, including a concise text description of the imagined information need, possibly augmented with video clips, still images and/or audio fragments that illustrate what type of video segments are needed. These topics express a wide variety of needs for such clips, e.g., showing particular objects ('sailing boat', 'pink flower'), people ('Ronald Reagan', 'Dr. Bondurant') and events ('water skiing', 'space shuttle landing').

The ground truth constructed in the evaluation process consists of the relevant video fragments, designated by their starting and ending times, where the evaluation assumes binary relevance. See (Smeaton, Over, & Taban, 2002; Smeaton et al., 2002) for further information on the Video Track framework.

The INEX document collection consists of the full texts of 12,107 articles from the IEEE Computer Society's publications between 1995-2002, totalling 494 megabytes in size and containing over eight million XML elements of varying granularity (from table entries to paragraphs, sub-sections, sections and articles, each representing a potential retrieval result). The topics of the test collection vary from natural language to structured texts in a modified XPath syntax. Based on the different topic types, INEX defined various ad-hoc retrieval tasks: content-only (CO), strict content-and-structure (S-CAS) and vague content-and-structure (V-CAS) retrieval. The S-CAS and V-CAS tasks are of no interest in this paper since in these tasks the target elements are defined (either strictly or vaguely) by the user. The CO task centres around the use of traditional IR-style queries that ignore the document structure. In this task, it is left to the retrieval system to identify the most appropriate granularity relevant XML elements to return to the user. See (Kazai et al., 2004) for more detailed information. For the construction of the relevance assessments, INEX employed two relevance dimensions, exhaustivity and specificity, each measured on multi-grade scales. A given component's degree of relevance, hence, combines a measure of how exhaustively it discusses the topic of request and a measure of how focused it is on the topic of request (i.e. discusses no other, irrelevant topics). The assessment procedure made explicit use of the nested XML structure to obtain assessments for each level of granularity, i.e. both ascendant and descendant elements of a relevant component had to be assessed. As a result, the ground-truth in INEX consists of nested relevant document components, i.e. subtrees, of the XML articles, where each such component is identified by its absolute XPath expression.

In general, both the corpus and the ground-truth of both collections can be simultaneously viewed as a hierarchy of nested components (i.e. frames, shots and scenes in video and nested elements in XML) and as continuous media (i.e. sequences of video frames and sequences of words/sentences). Given this parallel, the systems are assumed to retrieve relevant portions of the original documents, rather than the documents themselves, where the retrieved portions can be viewed as subtrees of the hierarchical structure or as parts of a video or text stream.

The remainder of this paper refers to these partial documents (i.e., video shots or XML document components) as *document fragments* (or fragments for short). The term *document* denotes either a video or an XML document. The primary goal of the paper is to show how to adapt IR evaluations for the case of retrieving fragments instead of documents.

3. Counting hits

Given a ranked list of retrieval results and a ground-truth set, the evaluation of IR systems is traditionally based on a mechanism that counts the number of retrieved documents that are also part of the relevance set. This method considers only exact matches, where the relevant reference item must be matched precisely by the retrieved item. Provided that a predefined unit of retrieval exists, such an evaluation procedure is adequate.

In the evaluation of video and XML retrieval systems however, the retrieved document fragments (*result fragments*) and the document fragments marked relevant in the ground-truth (*reference fragments*) are likely to have different starting points and lengths, hence, wide variations of overlap are possible between them. Given such a setting, a seemingly straightforward approach for evaluating a retrieved set of document fragments is to consider a result fragment f_i a correctly matched result if it intersects a relevant reference fragment f_r:

$$\text{match}(f_i, f_r) \iff f_i \cap f_r \neq \emptyset \tag{1}$$

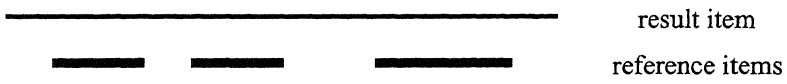

Figure 1. A single result overlapping multiple reference fragments

Unfortunately, this definition of a hit causes problems in the evaluation of the following two situations: (1) when a result fragment overlaps multiple relevant fragments and, (2) when multiple result fragments intersect the same relevant reference fragment.

3.1. Single Result Overlaps Multiple Reference Fragments

The first situation (visualised in Figure 1) encourages systems to return large document fragments, because doing so maximises the probability of a match. For example, a system that simply returns full documents in the collection as a result, would obtain almost perfect performance on recall and precision (only limited if the number of relevant documents in the collection would be larger than the number of results returned in the result set).

To resolve this issue, TRECVID 2001 introduced the following 'overlap measure' (Smeaton et al., 2002), which filters the result set on the length of the intersection between f_i and f_r:

KI coverage: The minimum value of the ratio of the length of the intersection to the length of the relevant fragment.

RI coverage: The minimum value of the ratio of the length of the intersection to the length of the result fragment.

In other words, KI coverage sets a lower limit for how much of the relevant fragment must be covered by the system's result, while RI coverage defines the lower limit for how much of the system's result fragment should overlap with the relevant reference fragment. In order to be considered a hit, a result fragment had to satisfy both constraints, where KI coverage was set to $\frac{1}{3}$ and RI coverage to $\frac{2}{3}$. The problem with this solution, however, is that current video retrieval systems have difficulty detecting semantic units of information in continuous media within a reasonable overlap of a human-specified result. Also, there is no clear motivation for a particular value of these thresholds, while the particular choice of a threshold affects the evaluation results significantly. Because no satisfying solution has been found, the next evaluations (TRECVID 2002 and 2003) have switched back to evaluation using predefined retrieval units.

INEX approached the problem of overlap differently, causing however increased complexity (and cost) during the assessments phase in creating the test collection. In this extensive assessment process, each ascendant and descendant component of a relevant element is assessed individually. During the evaluation, INEX then made use of two quantisation functions, strict and generalised, which provided a mapping of the two relevance dimensions of exhaustivity and specificity to a single relevance scale. In the strict case a binary relevance scale was applied and only the most exhaustive and most specific components were regarded as relevant. The generalised function used a 5-point linear relevance scale (with relevance values of 0, 0.25, 0.5, 0.75 and 1) and considered all components marked as relevant to some level.

In INEX, applying the quantisation functions on the exhaustive assessments eliminates the issue of matching multiple relevant components within a result, since only exact matches between result and reference fragments are considered as hits. For example, in the case of the strict quantisation, an XML result component that contains more than one relevant sub-components is only considered a hit if it matches

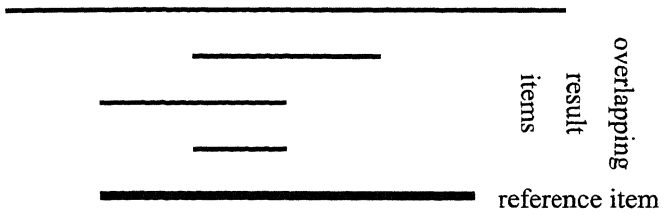

Figure 2. Multiple results overlapping a single reference fragment

exactly a relevant reference fragment (after quantisation). Similarly, in the generalised case, a result fragment is awarded the relevance value score associated with the reference fragment that it matches exactly (e.g. 0.75 if the reference component was judged highly exhaustive but only fairly specific).

3.2. Multiple Results Overlap Single Reference Fragment

The second situation, when multiple result fragments intersect a single reference fragment (see Figure 2), is problematic because matching based on intersection allows multiple system results to satisfy the same relevant fragment.

When computing recall and precision metrics, TRECVID's evaluation tool has treated this situation differently for recall than for precision (Smeaton et al., 2002). Recall was defined as the proportion of relevant fragments that has been found, hence counting each retrieved relevant reference fragment only once. But, precision was defined as the number of correct result fragments returned in the result set, therefore possibly counting the same relevant fragment multiple times. As a result, a system returning a series of N consecutive partially overlapping fragments, could get a precision at N of 100%, even if the topic has $R << N$ relevant fragments. Consider, for example, a topic with three relevant fragments r_1, r_2, and r_3 and two systems A and B, both returning 3 fragments. System A returns a list of result fragments that overlap with r_1, r_1 and r_2, while B returns r_1, r_2, and a non-relevant item. A and B perform equally on recall, but A is rewarded for returning r_1 twice, resulting in a precision of 100%; while system B has only a precision of 67%. Although system B's strategy requires less user effort to inspect the result list, its effectiveness is evaluated with the lower score.

Because the INEX evaluation framework takes the hierarchical structure of the document collection into account when making assessments, matching multiple results with one reference item is less of a problem (under the strict quantisation). Due to the XML structure, the only items that can overlap with a relevant reference item are the node itself, its ascendants, and its descendants. For the strict case, the ascendants and descendants of a relevant node are usually not relevant. When using generalised quantisation, Gövert et al. (Gövert, Kazai, Fuhr, & Lalmas, 2003) have proposed to resolve this problem by redefining precision to count each relevant fragment only the first time it is encountered as a match. This solution however relies on the assumption that relevant information is distributed uniformly throughout the component. Such an assumption is questionable whilst having a great impact on the evaluation.

So far in our discussions, we have highlighted the problems encountered when standard IR evaluation approaches are applied to video and XML retrieval, where the issue of handling overlap leads to questions on how to count hits. Although in some specific cases the counting problems are of limited impact, the overall problem remains unsolved. We base our solution for this counting problem on a model of the user performing the search task, described in the following Section.

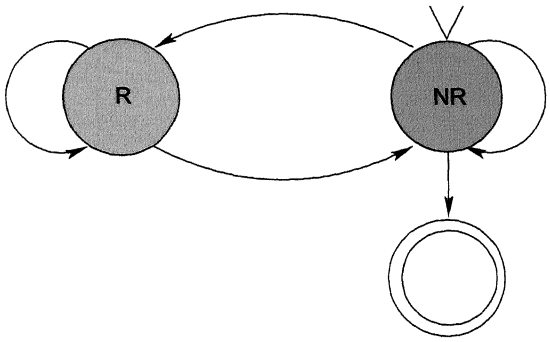

Figure 3. State Diagram of User Interaction.

4. Abstract User Model

Any measure to evaluate retrieval system performance takes assumptions about the expected user behaviour. Cooper stated that a retrieval system's main function is to 'save users work' (Cooper, 1968). This premise seems especially suited in the cases of video and XML retrieval, where we aim at finding the right fragments. Ideally, a retrieval system should show the user the relevant parts of the corpus only, without wasting user effort on the irrelevant parts. We propose to model this user effort by the *time spent on inspecting irrelevant information*. Section 5 develops from this model several metrics that estimate the duration that the user spends before having seen all relevant items, and variations that report the number of items found after, for example, one minute, or, conversely, the time spent before viewing s relevant items.

The underlying user model is based on the intuition that a user does not really care for an accurate *fragment* to be retrieved, but rather needs an entry-point into the document. Taking this view, a retrieval system can be considered to produce a ranked list of entry points. In the video case, the user starts viewing a video at a point where the systems suggests so. If the presented result is relevant, or it looks like relevance 'is starting to appear', he or she keeps viewing, and thus will see the relevant item if its starting point lies within a certain fixed window from the starting point returned by the system. In the XML case, the user starts reading the retrieved article from the suggested entry point, giving up when no relevant information is found for some time (or number of sentences). In other words, the user processes the retrieved information until his or her *tolerance to irrelevance* (T2I) has been reached, at which point the user proceeds to the next system result.

Tolerance to irrelevance is expressed by a single parameter for our user model, τ_{NR}. This parameter has a clear meaning in the real world, representing the maximum time that we expect a user would keep viewing non-relevant video, or, respectively, keep reading non-relevant text. The imaginary user views the video, or reads the XML document for τ_{NR} seconds, and if no relevant segment has been hit upon so far, he or she proceeds to the next retrieval result.

Figure 3 shows an abstract representation of such interaction between user and system. For each item, the user starts in non-relevant node NR. In the simplest case, the user switches automatically to relevant node R upon seeing relevant information (which could be immediate), and returns to the non-relevant state when there is no more relevant information present (i.e., the relevant item has ended). The non-relevant state can only be ended after wasting τ_{NR} user effort, or when more relevant information is encountered.

The simple model can be specialised to represent real-life more accurately. For example, a parameter τ_{init} could represent the initialisation time that the user needs to adjust to viewing a result item. Studies

with application of video tools in practise have shown that typically, τ_{init} would be about 3 seconds, while τ_{NR} varies between 10 and 15 seconds.* Dunlop based his work (with predefined retrieval unit) on a fixed time per web page (Dunlop, 1997). We are not aware of any existing experimental studies for the XML case, although work is underway, as INEX starts in 2004 an interactive track which aims to investigate how users work with XML IR systems. Here, a more complex user model might be needed to capture the user behaviour, e.g. when users navigate through the text using the tree structure instead of reading the text sequentially. Finally, notice that evaluation with a single value for τ_{NR} would assume that all users are the same. The fact that different users have different tolerances can be reflected by using a range of values in the evaluation.

As mentioned in Section 3, having no predefined retrieval unit leads to problems caused by multiple overlapping results and reference fragments. Adopting our user model discourages systems from returning fragments that are too large, since if the entry-point is too far away from the relevant reference fragment, the user's tolerance to irrelevance will have been exhausted before the relevant information has been reached. The problem with multiple system result fragments intersecting the same reference fragment is eliminated by extending the definition of irrelevance, according to which a previously seen reference fragment is no longer considered relevant.

The final issue to discuss is how to best handle the situation in which two relevant items appear close to each other in the corpus, i.e., when the gap between the two items is smaller than the threshold τ_{NR}. According to our user model, finding the first fragment automatically results in finding the second fragment if the gap between the two is sufficiently small. The user in fact might not even notice that there were two distinct relevant fragments. This raises the question whether we should simply merge these 'close' reference fragments together in the ground truth. Yet, if the system returns an entry-point in between the two results, then only one item is going to be viewed (assuming forward browsing only).

If we put the goal of understanding the merits of the distinct techniques applied in our system above the goal of estimating how appreciative an end user would be of our system, merging the reference items is undesirable. No matter how large τ_{NR} gets, we want to always distinguish between a system that identifies only the first and a system that identifies both items. For this reason, a binary property $\kappa_{\text{keepsviewing}}$ can be introduced in the user model, to decide whether a user continues viewing after the first relevant item has finished. If $\neg\kappa_{\text{keepsviewing}}$, the transition from node R to node NR is replaced by a transition from R directly to the end node.

5. Evaluation Metrics

The abstract user model specifies how to count hits from a system result list. We now derive T2I variants of existing evaluation metrics for system performance under a given instantiation of the abstract user model, i.e., when a suitable value for τ_{NR} has been chosen. Our T2I metrics are based on precision after a fixed amount of user-effort, the expected search length, and, the probability of relevant found. The common underlying principle is that retrieval systems are ranked on their ability to maximise the number of relevant fragments shown to the user while minimising the amount of user effort wasted on irrelevant information. The length of retrieved relevant fragments is ignored, assuming that each result has equal value to the user. Table 1 introduces our notation. For readability, we present the metrics only with user effort expressed as wasted time. Of course, in the XML case, the metrics can be defined more naturally by expressing the user effort in words or sentences instead.

Hull discusses in (Hull, 1993) the use of recall and precision at fixed document cut-off value (DCV), and points out that these measures normalise based on equivalent effort instead of equivalent performance. He suggests to measure precision over a range of DCVs and then average the results. To obtain a T2I variant of this metric, we will count the number of relevant fragments found before the available user effort has

*Personal communication related to (Amir et al., 2000).

Table 1. Notation.

Variable	Description
D	User effort required to inspect the full collection
R	Number of relevant fragments in the collection
D_R	User effort required to inspect all relevant fragments in the collection
j	Number of times T2I is reached while inspecting the result list
D_j	User effort wasted while inspecting result list
r	Number of relevant fragments not retrieved
s	Number of relevant fragments yet to be retrieved
i	Number of times T2I is reached while inspecting the non-retrieved corpus
I	Number of times T2I is reached when inspecting the full corpus
S	Number of relevant fragments desired

been wasted. The document cut-off value is defined in increments of τ_{NR}. If we let τ_{NR} correspond to 15 seconds, inspecting 20 results takes 5 minutes of the user's time (on the non-relevant information), so we expect a low document cut-off value (DCV) to be the more realistic choice. Under this condition, it is better to use precision than recall, because the number of relevant items might often exceed the DCV. To compute this measure, each time the user's tolerance to irrelevance has been reached, we measure the precision at that cut-off value. The process stops when we exhausted a predefined amount of time T, e.g., 5 minutes (for simplicity, take T a multiple of τ_{NR}):

$$\frac{\tau_{NR}}{T} \sum_{t=1}^{T/\tau_{NR}} \text{Precision after } t \cdot \tau_{NR} \text{ seconds wasted user effort} \tag{2}$$

When computing the precision after $t \cdot \tau_{NR}$ seconds wasted user effort, we take into account that multiple relevant fragments can be retrieved for each cut-off (i.e., when the gap between two relevant items is smaller than τ_{NR} and $\kappa_{\text{keepsviewing}}$). The metric obtained differs from Hull's proposal in that we exclude the relevant fragments from the user effort by which the effectiveness is normalised. Its main advantage is the intuitive interpretation as the average precision obtained at a fixed cost of wasted user effort. The obvious drawback of using just this metric is however that recall would be ignored completely.

The idea of measuring time passing as wasted user effort calls for Cooper's expected search length (ESL). ESL is defined as the expected number of irrelevant documents a user has to read to find a desired number of relevant items (S). Dunlop has used ESL in (Dunlop, 1997) to express user effort based on the time needed to inspect the result lists, which he called expected search duration (ESD), by assuming a uniform distribution of the lengths of retrieved items. Because his approach assumed a predefined retrieval unit, the expected search duration is then simply the multiplication of the expected number of documents retrieved with the average duration.

We adapt the Cooper proposal by replacing the original document retrieval model by our user model defining fragments from a user tolerance to irrelevance. In our setup, we cannot count the number of 'irrelevant fragments'. Like before, we choose to quantise the wasted user-effort as the number of times that the user's tolerance to irrelevance is reached. The ESL is then simply the user-effort wasted while inspecting the system's result list (j), augmented with the effort needed to find the remaining relevant items by random search through the collection. Cooper has shown that finding each of the s remaining relevant fragments requires inspection of an expected number of $\frac{i}{r+1}$ fragments, resulting in the following equation for ESL:

$$\text{ESL} = j + s\frac{i}{r+1}$$

In our case, the number of times the user threshold to irrelevance has been reached (i) cannot be counted directly. But, we can estimate its expected value from the known values of other variables. The cost of inspecting the non-retrieved non-relevant corpus equals $D - D_R - D_j$. Finally, we estimate the expected number of times that the T2I is reached by dividing this user effort in fragments of length τ_{NR}. We obtain the following closed-form formula to compute ESL:

$$\begin{aligned}
\text{ESL} &= j + s\frac{i}{r+1} \\
&= j + s\frac{1}{r+1}\left\lceil\frac{D - D_R - D_j}{\tau_{NR}}\right\rceil \\
&= j + s\frac{1}{r+1}\left\lceil\frac{D - D_R - j \cdot \tau_{NR}}{\tau_{NR}}\right\rceil \\
&= j\frac{r+1}{r+1} + s\frac{1}{r+1}\left\lceil\frac{D - D_R}{\tau_{NR}}\right\rceil - \frac{j \cdot s}{r+1} \\
&= j\frac{r-s+1}{r+1} + s\frac{I}{r+1}
\end{aligned} \tag{3}$$

In the final step, we use the fact that $I = \left\lceil\frac{D-D_R}{\tau_{NR}}\right\rceil$. The result of Equation 3 can be normalised by comparing it to the expected random search time, i.e. the time the user would have spent to find *all* items without a retrieval system ($S\frac{I}{R+1}$). If we subtract the result from 1, we get Cooper's expected search length reduction factor, which expresses how much better the retrieval system works than searching purely at random:

$$\begin{aligned}
\text{ESLRF} &= 1 - j\frac{r-s+1}{r+1}\frac{R+1}{S \cdot I} - s\frac{I}{r+1}\frac{R+1}{S \cdot I} \\
&= 1 - \frac{R+1}{S(r+1)} \cdot (s + \frac{j(r-s+1)}{I})
\end{aligned} \tag{4}$$

Finally, Raghavan et al. have shown in (Raghavan, Bollmann, & Jung, 1989) that the probability of relevance $P(Rel|Retr)$ for R relevant fragments retrieved can be computed from the expected search length as follows:

$$P_R(Rel|Retr) = \frac{R}{R + \text{ESL}_R} \tag{5}$$

The advantage of this latter metric, which is used at INEX, is that it is theoretically justified to choose its computation at an arbitrary recall point $x \cdot R$ as well (with $x \in [0, 1]$). This allows the evaluation of retrieval systems using multiple queries, by averaging the scores for each query at the same set of recall points.

We have shown how existing evaluation metrics can be adapted to our approach, based on user-effort and the notion of tolerance to irrelevance. Equation 2 gives the T2I variant of Hull's proposal to measure precision averaged over multiple document cut-off values. It provides a simple and intuitive characteristic of a retrieval system, that is easily interpretable. Equation 3 gives a T2I variant of the closed form formula for computation of ESL. A complete recall-precision graph over a fixed set of recall points can be constructed by using the probability of relevance by Raghavan et al., given in Equation 5.

So far, the DCV variant of the metric has been implemented as an experimental extension of the NIST evaluation tool that was used at TRECVID 2001. We are currently looking into implementation of the Raghavan variant of our proposed metric by extending the INEX evaluation tool.

6. Conclusions

Both video and XML retrieval systems return fragments instead of full documents, posing a problem for the standard evaluation approaches in information retrieval, which assume a known, predefined retrieval unit. The difficulty lies in the question of how to count the number of relevant fragments retrieved if a returned system item overlaps with multiple reference items, or, conversely, if one reference item intersects multiple system results.

The first contribution of this paper is a detailed analysis of these problems with IR system evaluation without predefined retrieval unit. Next, we approached the problem by focusing the evaluation on user-effort. In the case of retrieval without predefined retrieval unit, minimising user-effort is an important part of the search task, in that we do not want systems to retrieve too large fragments that contain only a minimal amount of relevant information. Also, systems should be discouraged from returning multiple results that overlap with each other.

We based our proposed evaluation measures on an abstract model of the user, characterised by a single parameter that models the user's tolerance to irrelevance. We demonstrate how this model can be employed within existing metrics emphasising user effort, such as (Hull, 1993; Cooper, 1968; Raghavan et al., 1989). Our approach is similar in spirit with Dunlop (Dunlop, 1997), but focuses on retrieval of document fragments rather than full documents only. The resulting metrics form attractive alternatives to the existing approaches of assessing system performance.

We believe our approach results in a simple and effective solution to the problems in the evaluation of retrieval systems without predefined retrieval unit. The results apply to the two evaluations that have been discussed in detail, but can also be used for other retrieval problems like passage retrieval and spoken document retrieval.

The next step in our research is to gain experience from using the metrics in practice for the evaluation of retrieval systems. We need experimental studies, to calibrate the τ_{NR} parameter, and, to establish the stability of system rankings with respect to variations in this parameter, like e.g. (Buckley & Voorhees, 2000). An interesting direction for further research is the usage of alternative stopping criteria in the ESL measure, like in (Kraft & Lee, 1979). Also, we could differentiate between the cases when a system returns results that are not relevant to the search topic and when it returns results that have already been seen. The user may tolerate redundancy in results more than the wasted effort of viewing irrelevant results, which should be reflected in the evaluation metrics. Finally, we plan to investigate how the current proposal combines with evaluation using the cumulative gain metrics of (Järvelin & Kekäläinen, 2002), to incorporate the graded assessments used at INEX.

Acknowledgements

The authors are grateful to the IBM Research team that participated in TRECVID 2001 (especially Savitha Srinivasan) for discussions regarding the analysis of the obtained retrieval results.

References

Amir, A., Ponceleon, D., Blanchard, B., Petkovic, D., Srinivasan, S., & Cohen, G. (2000, Jan). Using Audio Time Scale Modification for Video Browsing. In *Proceedings of the Thirty-Third Hawaii Int. Conf. on System Sciences, HICSS-33*. Maui, Hawaii: Computer Society Press.

Buckley, C., & Voorhees, E. M. (2000). Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 33–40). Athens, Greece: ACM Press.

Cooper, W. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation, 19*(1), 30–41.

Dunlop, M. (1997). Time, Relevance and Interaction Modelling for Information Retrieval. In *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)* (pp. 206–213). Philadelphia, PA, USA: ACM Press.

Gövert, N., Kazai, G., Fuhr, N., & Lalmas, M. (2003). *Evaluating the effectiveness of content-oriented XML retrieval* (Technischer Bericht). Dortmund, Germany: University of Dortmund, Computer Science 6.

Harman, D. (Ed.). (1992). *Proceedings of the First Text Retrieval Conference (TREC-1)* (Nos. 500–207).

Hull, D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)* (pp. 329–338). Pittsburgh, PA, USA: ACM Press.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems (ACM TOIS)*, *20*(4), 422–446.

Kazai, G., Lalmas, M., Fuhr, N., & Gövert, N. (2004). A Report on the First Year of the INitiative for the Evaluation of XML Retrieval (INEX'02), European Research Letter. *Journal of the American Society for Information Science and Technology (JASIST)*. (To appear)

Kraft, D., & Lee, T. (1979). Stopping Rules and their Effect on Expected Search Length. *Information Processing and Management*, *15*, 47–58.

Raghavan, V., Bollmann, P., & Jung, G. (1989). A Critical Investigation of Recall and Precision. *ACM Transactions on Information Systems (TOIS)*, *7*(3), 205–229.

Smeaton, A., Over, P., Costello, C., Vries, A. de, Doermann, D., Hauptmann, A., Rorvig, M., Smith, J., & Wu, L. (2002, September). The TREC2001 Video Track: Information Retrieval on Digital Video Information. In *Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002, Proceedings* (Vol. 2458, pp. 266–275). Rome, Italy: Springer.

Smeaton, A., Over, P., & Taban, R. (2002). The TREC-2001 Video Track Report. In E. Voorhees & D. Harman (Eds.), *Proceedings of the Tenth Text Retrieval Conference TREC-10* (pp. 52–60). Gaithersburg, Maryland, USA: Department of Commerce, National Institute of Standards and Technology.

Tague-Sutcliffe, J. (1992). The Pragmatics of Information Retrieval Experimentation Revisited. *Information Processing and Management*, *28*(4), 467-490.