Inf Retrieval (2010) 13:460–484
DOI 10.1007/s10791-010-9133-9

Expected reading effort in focused retrieval evaluation

Paavo Arvola · Jaana Kekäläinen · Marko Junkkari

Received: 1 May 2009 / Accepted: 14 April 2010 / Published online: 6 May 2010

Abstract This study introduces a novel framework for evaluating passage and XML retrieval. The framework focuses on a user's effort to localize relevant content in a result document. Measuring the effort is based on a system guided reading order of documents. The effort is calculated as the quantity of text the user is expected to browse through. More specifically, this study seeks evaluation metrics for retrieval methods following a specific fetch and browse approach, where in the fetch phase documents are ranked in decreasing order according to their document score, like in document retrieval. In the browse phase, for each retrieved document, a set of non-overlapping passages representing the relevant text within the document is retrieved. In other words, the passages of the document are re-organized, so that the best matching passages are read first in sequential order. We introduce an application scenario motivating the framework, and propose sample metrics based on the framework. These metrics give a basis for the comparison of effectiveness between traditional document retrieval and passage/XML retrieval and illuminate the benefit of passage/XML retrieval.

Keywords Passage retrieval · XML retrieval · Evaluation · Metrics · Small screen devices

1 Introduction

The traditional information retrieval (IR) considers a document to be an atomic retrievable unit. Since not all content of a document is relevant according to a query, it is useful to retrieve smaller parts e.g. with an XML retrieval system or a system retrieving arbitrary passages. This enables a more specific retrieval strategy and allows a system to focus on

P. Arvola (✉) · J. Kekäläinen
Department of Information Studies and Interactive Media, University of Tampere, Tampere, Finland
e-mail: paavo.arvola@uta.fi

J. Kekäläinen
e-mail: jaana.kekalainen@uta.fi

M. Junkkari
Department of Computer Sciences, University of Tampere, Tampere, Finland
e-mail: junken@cs.uta.fi

parts of documents. Thus, content-oriented XML retrieval and passage retrieval are beneficial in reducing a user's effort in finding the best parts of a document.

From the evaluation perspective, the fundamental difference between content-oriented XML retrieval and passage retrieval is that in XML retrieval the passages are marked-up as elements, i.e. text is between the element's start and end tags, whereas in passage retrieval the passages are not dependent on element boundaries. In this study the term passage retrieval is extended to concern content-oriented XML retrieval as well.

The retrieved passages can be grouped in many ways. This study follows a specific fetch and browse approach (Chiaramella et al. 1996). In the fetch phase documents are ranked in decreasing order according to their document score, just like in the traditional document retrieval. In the browse phase, a set of non-overlapping passages representing the relevant text within a document is retrieved, and the retrieval system interface turns the searcher's attention to the relevant parts of the documents. The matching method, including the selection of appropriate (best matching) passages, defines *what* the user is expected to browse. The user interface, in turn, specifies *how* the user is expected to browse the content. The co-operative action affects the reading order of the passages, and the effort the user has to spend in localizing the relevant content in the document. This effort can be measured by the quantity of text the user is expected to browse through.

The amount of text can be measured e.g. with words, sentences or windows of characters. We have chosen to use characters as the measurable units of the user effort. Characters are the smallest atomic units of text to read, retrieve and evaluate, and we assume a character to be read with a constant effort. This is tolerable while by comparison in document retrieval the effort of reading a whole document is treated as a constant regardless of the size of the document and other qualities.

Considering characters to be retrievable units, any text document can be modeled as a character position list, starting basically from the first character (at position 1) and ending at the last character (at position n) of the document (n is the length of the document). Other characters are in sequential order in between.

This order can typically be considered as the author's intended reading order of the document. In other words, the author 'expects' the reader to follow the sequential order of the document. Thus, the corresponding character position list for the intended reading order is $\langle 1,2,3,\ldots,n \rangle$. Figure 1 illustrates this.

However, in reality this order applies only for a person who is a strict sequential reader; other kinds of readers probably follow some other order of their own, based on numerous uncontrollable factors beyond the author's intention (see Hyönä and Nurminen 2006). Yet, any reading order can also be modeled as a character position list, but the order of the list items follows rather a temporal pattern of the reader's behavior and the list may differ more or less from the intended reading order character position list.

The user behavior is not a totally independent variable. Namely, the usage of a passage retrieval system in co-operation with a user interface provides means to guide browsing within a document, and to break the intended reading order and re-organize the expected browsing order. This kind of system guided reading order is supposed to allow more focused access to the relevant content in a document. Thus, the aim of this study is to provide an evaluation framework based on the guided reading order within a document.

In addition to the expected reading order, one might expect that the user is not willing to browse through all irrelevant material, if a lot of such material is met. Instead, the user will stop at some point and move onto the next document (or reformulate the query). This means that not all relevant content of a document, if any, may be encountered in the event the user has to read a lot of irrelevant material.

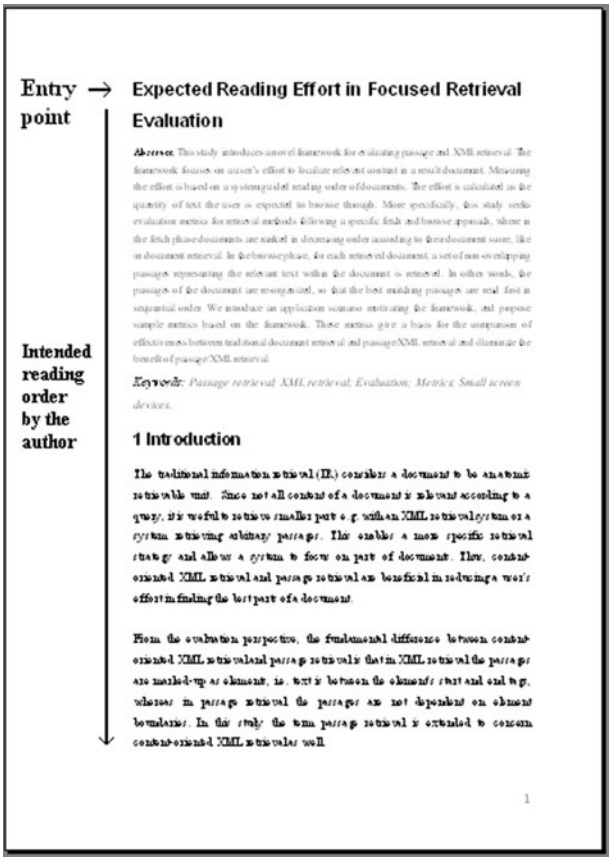

Fig. 1 Intended reading order of a document by the author (Greek and Latin based scripts)

The benefit of the evaluation framework is twofold: First, since the user effort is based on characters to be read, it is credible and adjustable for various user interfaces and user scenarios. Second, it gives a basis for effectiveness comparison between traditional document retrieval and passage/XML retrieval. The latter illuminates the benefit of passage/XML retrieval, which has been questioned by Kamps et al. (2008a).

For an IR system's tasks in browsing a document, we propose two approaches: (1) to quickly assess the document to be relevant (or not relevant) and (2) to browse through the relevant content of a document effectively. For the two tasks we introduce two metrics in Sect. 4, namely *cumulated effort* for the first and *character precision-recall* for the other.

To motivate and clarify the evaluation framework, we present a sample user interface scenario in Sect. 2. However, it is worth noting that the presented framework is independent of the sample scenario. Section 3 reviews related studies. In Sect. 4 we introduce two metrics with sample measures, and test results on Wikipedia data in Sect. 5.

2 Motivating sample scenario

The reading order of a document depends on the co-operative action of the user interface and the passage matching method. Therefore, as a sample scenario and a basis of evaluations, we introduce an interface, which is a slightly simplified variant of the interface

described by Arvola et al. (2006). This scenario motivates the present study and is the basis of the experiments, but the evaluation framework is not bound to any of its features (incl. screen size or other technical details). It is necessary to emphasize that this sample scenario is only illuminating.

Passage and element retrieval provide focused access to documents. The size of passages/elements may vary but they are always accessed through a window, size of which depends on the media and device. This feature is stressed when using a device with limited screen space, such as a mobile phone. The small screen forces the user's attention to the position the screen is showing, and thus the expected user behavior is more predictable.

A small screen is one of the major constraints for a mobile device. Because of that, several approaches in preventing horizontal scrolling are introduced (Buyukkokten et al. 2000; Jones et al. 1999). Our sample scenario follows the Opera Mini browser (Opera 2006) outline, where the textual content of a document is rendered in one column. Nevertheless, conventional browsing through a long text document with such a device requires a lot of vertical scrolling.

In our interface the effort of finding relevant content is reduced by inserting hyperlinks into anchors that, in turn, are placed at the matching locations of the document according to the initial query expression. The user is directed to the supposedly relevant parts of the document.

This user interface not only reduces the user's effort in reduced vertical scrolling but also preserves the original document order and the structure of the document. In other words the document is represented without breaking up the continuity of the initial textual content presentation of the document. The conventional browsing methods within the document are also available. Consequently, the user is expected to navigate through the document in the fashion described in Sects. 2.1 and 2.2.

The anchoring of passages is done simultaneously with rendering documents to the standard XHTML format for viewing and browsing in a device independent way. In XML retrieval, our method is especially suited for content-oriented online XML collections, such as the Wikipedia XML collection used in INEX 2008 (Denoyer and Gallinari 2006). Next, we give a detailed view of the system.

2.1 Interface overview

In a search process the user inserts keywords with available text input methods in order to perform a search. In order to retrieve the best matching passages from the best matching documents, the retrieval follows the fetch and browse approach. In a nutshell, documents are first sorted according to their retrieval status value, and after that the passages are clustered by documents. Phase 1 is plain full document retrieval where, according to the query expression, the system presents a result list with links to the documents in a matching order. Figure 2 illustrates this phase. Phase 2 is element retrieval and it is done for a single document selected by the user.

The usage of the interface follows strictly the fetch and browse approach. In Fig. 2 the query: 'matching method' is forwarded into the IR system. This launches the full document retrieval phase and the system presents a result list, with the current document on top of it. Preferably, the user selects this document from the result list by clicking a pointing link. Clicking the link triggers Phase 2 (Fig. 3). In this phase, the system marks up the best matching parts of the document. Thereafter the system renders the resulting document into an XHTML document for viewing and browsing. This includes also inserting anchors into the beginning of the best matching parts. Finally, the browser shows the beginning of the result document to the user.

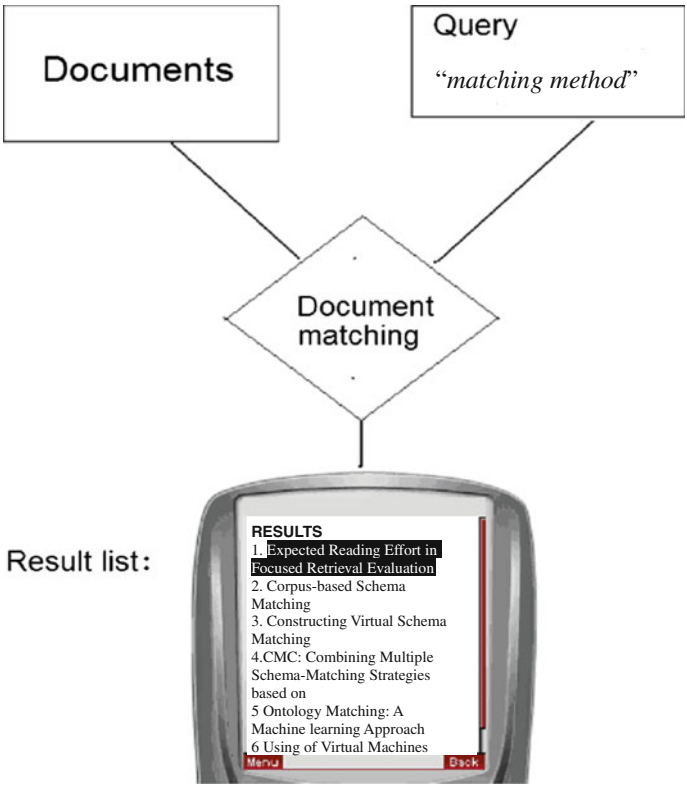

Fig. 2 Phase 1 (fetch)—document ranking

2.2 Creating a matching chain by linking the Best parts with anchors

In the present user interface scenario, there are two arrow icons at the beginning of the resulting document. The first one, an arrow down, is a link to the first anchor at the beginning of the first matching element. The arrow left is a link to getting back to the result list. The arrow down hyperlink is for relevance browsing within the document. By clicking it, the user ends up to the point the anchor is at. At the end of the matching passage, the arrows are presented again. Now the arrow down is a link to the next matching (not overlapping) part of the document.

For instance in Fig. 3, the user selects the current document from the result list. The system places two arrow hyperlinks to the top left corner of the result document. In Fig. 4 the current user interface focus is on the first hyperlink, which is represented by an arrow down hyperlink. By clicking the link the user moves down to the place the anchor is at. If the matching works perfectly the anchor is just before a relevant part in the result document. Now the user scrolls and reads the whole section, which is estimated to be relevant by the retrieval system. Because there are no further relevant parts in the document at hand, at the end of the section there are only two hyperlinks: back to results and back to the beginning of the document. In case there were other relevant passages further down in the document, there would be an arrow down hyperlink to the start of the next matching passage and so on.

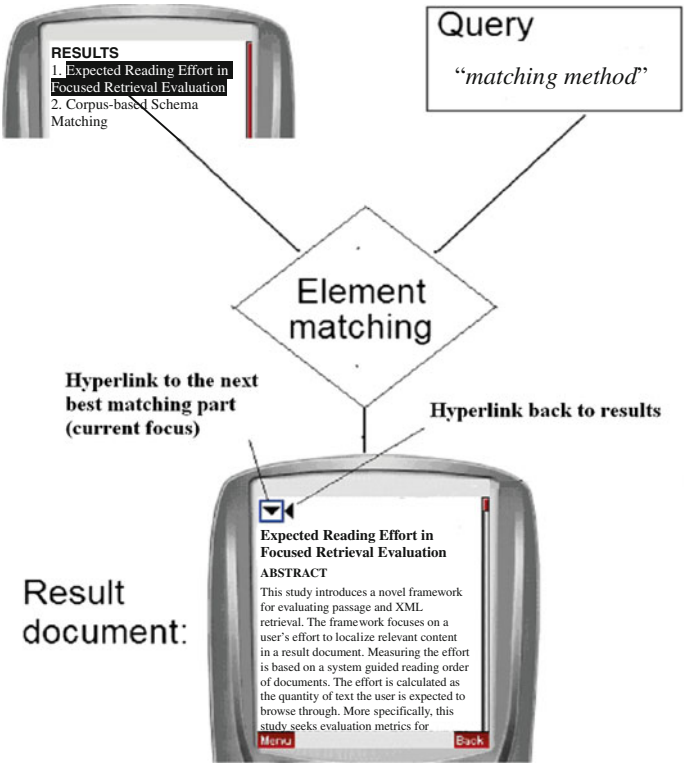

Fig. 3 Phase 2 (browse)—relevant in document retrieval

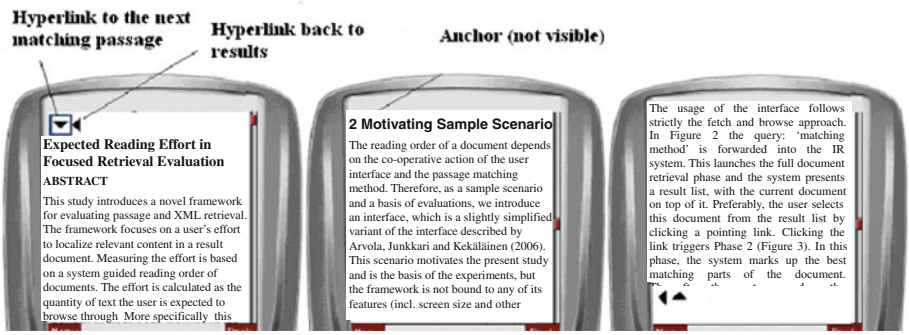

Fig. 4 Matching chain

When a user has read the retrieved passages and is still willing to read more within the document, we have to consider options of how the user proceeds after seeing the last retrieved passage. The extreme options for further reading are that the remaining relevant text is read immediately after the last retrieved passage (best case), or after all other non-relevant text is read (worst case). The best case can be discarded, since it is too easy to deliver good results with that. Instead, we define a third case in between (natural case),

where the reader clicks the hyperlink back to the beginning of the document and reads the remaining parts in document order.

The scenario affects the reading order but it does not tell us how long the user is willing to browse the document. We assume that the browsing continues until the user's *tolerance to irrelevance* (de Vries et al. 2004) has been reached. At that cut-off point the reader is assumed to be bored with the irrelevant material and moves onto the next document in the result list.

Consequently, evaluating the results delivered by a retrieval system can be based on this scenario. In other words, the user is expected to follow the matching chain, i.e. read the matching passages in document order. Whether the interface is easy to use, or whether every user is willing to utilize the features in this scenario are important usability issues but not a concern of the present study. In Sect. 5 we illustrate, with the metrics described in Sect. 4, how different retrieval systems perform within this sample scenario and how systems perform without the anchor hyperlink structure (i.e. document retrieval). In other words, what is the improvement rate, when using passage retrieval and the presented user interface compared with traditional document retrieval?

3 Related work

The present study combines measuring passage/XML retrieval and the concept of expected user effort. Accordingly, in Sect. 3.1 we review the current metrics evaluating fetch and browse style XML/passage retrieval and in Sect. 3.2 we address existing approaches in measuring the user effort at the retrieval.

3.1 Passage retrieval and relevant-in-context in INEX

The evaluation framework presented in this study is focused on a reading order within a single document and the approach is close to passage or XML retrieval and the work done within this context. Therefore the outcome of the present study relates to INEX (Initiative of the Evaluation of XML, INEX 2009) and the evaluation testbed provided by it. INEX is a prominent forum for the evaluation of XML retrieval offering a test collection with topics and corresponding relevance assessments, as well as various evaluation metrics. Currently, aside evaluating element retrieval, passage retrieval evaluation is also supported in INEX. That is because the relevance assessments are executed so that the assessors have marked up the relevant passages regardless of any element boundaries (Piwowarski and Lalmas 2004). Similar relevance assessments are available also in TREC Hard Track's passage retrieval (Allan 2004). In terms of the present study, a recall base for a document consists of a set of character positions.

The evaluation of the fetch and browse approach is an essential issue in content-oriented XML retrieval. This is mainly because the Relevant-in-Context (RiC) task of the INEX's tasks is considered the most credible from the users' perspective (Trotman et al. 2007; Tombros et al. 2005). The task corresponds fully to the fetch and browse approach. Because of the complex nature of the task, the evaluation measures are constantly evolving. There has also been a concern that full document retrieval would be very competitive in XML retrieval (Kamps et al. 2008a). According to the fetch and browse approach, in the official metric of the RiC task, separate scores are calculated for each individual retrieved document d as a *document score* ($S(d)$) in the browse part, and the

document result list as a *list score* in the fetch part. Next we introduce the current official metric for RiC in detail.

3.1.1 List score

The list score is calculated over a ranked list of documents based on document scores. A generalized precision is calculated as the sum of document scores up to an article-rank divided by the article-rank. Similarly generalized recall is the number of relevant articles retrieved up to an article rank, divided by the total number of relevant articles. Formally generalized precision (gP) at rank r is defined as follows:

$$gP[r] = \frac{\sum_{i=1}^{r} S(d_i)}{r},$$

and similarly the generalized recall:

$$gR[r] = \frac{\sum_{i=1}^{r} isrel(d_i)}{Trel},$$

where *Trel* denotes the total number of relevant documents and *isrel* is a binary function of the relevance at a given point. With these equations, we are able to calculate the average generalized precision for the result list:

$$AgP = \frac{\sum_{r=1}^{D} (isrel(d_r) \times gP[r])}{Trel},$$

where D is the ranked list of documents. Mean average generalized precision ($MAgP$) is calculated basically as the mean of the values of individual topics. Further details can be found in (Kekäläinen and Järvelin 2002; Kamps et al. 2007, 2008b, c).

The list score is general in a sense that the calculation of document score ($S(d)$) is not predefined, except that the values range is [0,1]. We adopt the list score for our evaluation metrics and replace later the document score with our own formula. Next, we introduce the official document measure used in INEX.

3.1.2 Document score in INEX

The official INEX measure for the document score is an *F-Score* of the retrieved set of character positions (Kamps et al. 2008b; see also Allan 2004). The *F-Score* is calculated for each retrieved document d as follows (Kamps et al. 2008a):

$$F_\alpha(d) = \frac{(1+\alpha^2) \times P(d) \times R(d)}{\alpha^2 \times P(d) + R(d)},$$

The α value is used to tune the role of the precision in the formula. It determines the power of which the precision is taken into account in the evaluations. $P(d)$ (the document precision) is the number of retrieved relevant characters divided by the number of retrieved characters. $R(d)$ (the document recall), accordingly, is the number of characters assessed to be relevant that is retrieved divided by the total number of relevant characters as follows:

$$P(d) = \frac{|rel(d) \cap ret(d)|}{|ret(d)|}$$

$$R(d) = \frac{|rel(d) \cap ret(d)|}{|rel(d)|}$$

In other words, the retrieval performance of a system is based solely on the set of character positions within the retrieved passages, whereas our approach takes the reading order and the tolerance of irrelevance into account as well.

The aim of using the *F-Score* of retrieved passages is to measure effectiveness in the relevant in context task, where "focused retrieval answers are grouped per document, in their original document order, providing access through further navigational means" (Kamps et al. 2008b). However, since the official *F-Score* measure treats the retrieved character positions as a set, the reading order dimension remains unjustified. With the framework of this study we take the document order (i.e. reading order) into account. Based on the framework we present novel measures as alternatives to the *F-Score* of retrieved passages in Sect. 4.

Since the final score of a system is a combination of document and list scores, we denote the combined measure as *list_score\document_score*. For example the official INEX measure, mean average generalized precision list score over $F_{0.25}$ document scores is denoted as $MAgP\backslash F_{0.25}$.

3.2 User effort in evaluation metrics

Most evaluation metrics of IR effectiveness are based on topical relevance (Saracevic 1996) and include explicit or implicit user models. Besides relevance, evaluation metrics have tried to encompass other aspects of user behavior affecting information retrieval, most prominently satisfaction and effort. Next we review metrics related to our study.

The implicit user model of the traditional (laboratory) full document retrieval evaluation assumes that the user reads the documents of the result list one by one starting from the beginning, and stopping when the last relevant document is passed. This might not be a realistic assumption of the user behavior, but has been considered adequate for evaluation purposes for decades. A further elaboration (Robertson 2008) interprets *average precision* with an assumption that users stop at a relevant document in the ranked list after their information need is satisfied. If stopping is uniformly distributed across relevant documents, average precision may be interpreted according to this simple user model.

Expected search length (ESL, Cooper 1968) takes the expected user effort into account as the average number of documents the user has to browse in order to retrieve a given number of relevant documents. ESL has inspired other metrics, like *expected search duration* (Dunlop 1997) and *tolerance to irrelevance* (T2I, de Vries et al. 2004). Instead of search length, expected search duration measures the time that users need to view documents in the ranked result list to find the required number of relevant documents. Predicted user effort is incorporated with the interface and search engine effects into one evaluation model.

A user's tolerance to irrelevant information is a central notion in T2I. It is aimed at retrieval environments without predefined retrieval units; in other words, passages of documents (or other information storage units) are retrieved instead of whole documents. The user model of T2I assumes that a retrieved passage acts as an entry point to the document: if the user does not find relevant information in the document before his or her tolerance to irrelevance is reached, he or she will move to the next item in the result list. de Vries et al. (2004) combine user effort with the model, and propose measuring it as time spent on inspecting irrelevant information. Moreover, they mention that in XML IR words or sentences could be used as well. As actual evaluation measures the authors propose a T2I variant of average precision of document cut-off values, i.e. average over precisions after

given T2I points in time, and an ESL variant, i.e. "the user effort wasted while inspecting the system's result list … augmented with the effort needed to find the remaining relevant items by random search through the collection" (de Vries et al. 2004, 470).

Other metrics combining user effort with retrieval evaluation are *expected precision-recall with user modelling* (EPRUM, Piwowarski and Dupret 2006; Piwowarski 2006) and *effort precision—generalized recall* (EP/GR, Kazai and Lalmas 2006). EPRUM considers returned result items as entry points to the collection from which points the user can navigate to relevant items. EP/GR measures the amount of effort as the number of visited ranks that the user has to spend when browsing a system's ranked result list, compared to the effort an ideal ranking would take in order to reach a given level of gain.

Like T2I our framework can be applied not only to XML elements but also to arbitrary passages. Our framework shares the notion of effort with the earlier measures; however we interpret effort at character level but in principle share the idea of measuring effort as the amount of text. Further, the best retrieved passage acts as an entry point to the document like in T2I and EPRUM. We also utilize tolerance to irrelevance as a stopping rule within a document. However, our framework exploits the guided reading order of the retrieved document and scoring of each document is based on this order. Consequently measuring effectiveness at document level differs from other measures.

In our framework, the reading order is considered within a document, not in a hierarchical element structure. The idea behind the browsing model is somewhat different than in EPRUM (and also Ali et al. 2008), where the browsing is based on hierarchical and linked items. The proposed framework differs from the earlier work by combining the character level evaluation to the system guided reading order.

4 Metrics based on expected browsing effort

The character level evaluation can be associated with the traditional document retrieval evaluation in a sense that the retrievable unit is a character and is treated as if it was a document. When considering the result as a set of retrieved character positions, document's precision and recall and document's *F-Score* correspond to the full match evaluation measures. Our approach, however, resembles partial match evaluation on the document level. Instead of treating the result as a set of character positions, we treat the result as a list of character positions. The order of the characters in the list depends on their browsing order. Clearly, treating the retrievable units as a list instead of a set broadens the number of alternatives for the retrieval performance measures. In addition, the list approach brings on the temporal dimension in browsing, and thus enables the exploitation of the T2I approach.

We present two metrics based on the reading order: *character precision-recall* (*ChPR*) and *cumulated effort* (*CE*). For both metrics we assume that some text within the retrieved relevant document is assessed to be relevant. In other words there exists a recall base, like the INEX recall base, containing the character positions of relevant characters. These characters are then compared with the expected order of reading. The metrics follow the underlying evaluation framework and the reading order is not bound to any specific user or interface scenario.

In *ChPR* the relevance score values scale between 0 and 1, and the list score is calculated analogously to generalized precision-recall, whereas in *CE* the document scoring is looser and the list score is calculated by cumulated effort, which has evolved from the cumulated gain metric (Järvelin and Kekäläinen 2002).

4.1 Character precision-recall

In our framework characters are units to retrieve and they are expected to be read in some order. This simple reading model along with the character position wise relevance assessments enables the usage of the standard precision-recall metric for the document score, and thus all the related measures are available. The list score, in turn, is calculated by the generalized precision-recall metric as given in Sect. 3 (Kamps et al. 2007).

As a basic measure of this metric, we define the character average precision for a document, *aveChP*(*d*), which is similar to the traditional average precision measure. The only difference is that documents are replaced here with characters.

$$aveChP(d) = \frac{\sum_{p=1}^{|d|} (P_d(p) \times RL_d(p))}{NRC_d}$$

In the formula, p is the character position from the point the reading starts, *RL* a binary valued function on the relevance of a given position, *NRC* the number of relevant characters in document d, and P precision at a given position in d. We set *aveChP* as the document score for calculating the list score with the generalized precision-recall metric (see Sect. 3.1.1). Note that *aveChP* is calculated for a relevant document only. For non-relevant documents *aveChP*, and other measures within the character precision-recall metric the document score is 0.

The *aveChP* can be considered a somewhat system-oriented measure, since it does not take a stand on when the user stops reading the document. However, it rewards systems that organize the expected reading order in an optimal way. Naturally, instead of using *aveChP* a number of cut-off measures can be used, for instance precision can be calculated when a chunk of 600 characters is read (i.e. *ChPR*@600). Apart from this kind of basic cut-off point, a user oriented cut-off point, like T2I, can be utilized. In T2I the reading of a document is supposed to end when a pre-set tolerance to irrelevance has been reached (or the whole document is read through). For instance the T2I measure *T2Iprec* (2000) means that the reading ends, when the user has seen 2000 non-relevant characters, and then document's precision is calculated. In other words the T2I is a cut-off measure, where the cut-off point varies according to the read irrelevant material. In addition to precision, also recall (*T2Irecall*) and their harmonic mean *F-Score* ($T2If_\alpha$) are viable measures to combine with T2I. Note that this time the *F-Score* is calculated according to the read characters, not to the retrieved.

A couple of toy examples illustrate *aveChP*; we calculate some sample values for a 'mini document'. In the example the reading order is based on the scenario given in Sect. 2 and the natural case reading order for the non-retrieved passages. In Table 1, there is a character position list for a sample mini document.

For each example the characters assessed as relevant are in bold face and the retrieved characters are underlined. The two examples are the following:

Example 1: "**relevant content is in bold** and retrieved is underlined"
Example 2: "**relevant content is in bold** and retrieved is underlined"

Example 1: The system has found a relevant document (value as such), but is unable to identify the relevant content within the document. The expected reading order is ⟨33, …, 55, 1, 2, …, 31, 32⟩ and the recall base is the set {1, 2, …, 27}. Thus $aveChP = 0.35$. The *F-Score* (of the retrieved characters, $\alpha = 1$) does not give any value to this result, and thus the document corresponds to a non-relevant document: i.e. $F\text{-}Score = 0$. However, in this case the passage retrieval system is not helpful, because the relevant content is at the

Table 1 Character position list of a mini document (line break is nr. 28)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
r	e	l	e	v	a	n	t		c	o	n	t	e	n	t		i	s		i	n		b	o	l	d
29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55
a	n	d		r	e	t	r	i	e	v	e	d		i	s		u	n	d	e	r	l	i	n	e	d

beginning of the document and no guidance is needed. Thus, a system retrieving full documents would deliver the following scores: $aveChP = 1$, $F\text{-}Score = 0.66$.

Example 2: The passage retrieval system does not identify all relevant content, but the retrieved and relevant passages are partially overlapping. This example illustrates the early precision biased nature of the *aveChP* measure. Here, *F-Score* is 0.16. The reading order is: ⟨24, 25, …, 44, 45, 1, 2,…, 23, 46, …, 55⟩, and $aveChP = 0.53$.

Comparing our reading order based approach to the *F-Score* shows the benefit of combining the amount of read text and reading order in evaluations. For example *F-Score*, would give the same score to a long document, with a relevant latter half, as with a relevant first half, even though it requires more effort to reach the latter half, assuming that the browsing starts at the beginning of the matching passage, in this case the whole document.

In addition, the *F-Score* calculation involves a hidden assumption that the user stops reading the document after the retrieved passages. This holds even if there were still relevant passages elsewhere in the document to be read. Thus, the browse phase for reaching more relevant content is incomplete, and the passages of the next document in the result list are prioritized over the ones in the document the user is currently looking at. These passages will then never be reached. This seems a rather simple user model.

4.2 Cumulated effort

Instead of the gain the user receives by reading the documents in the result list, cumulated effort (*CE*) focuses on the effort the user has to spend while looking for relevant content. The effort-oriented metric should fulfill the following aims: (1) to model the increase of the expected effort, when the user is reading the document list further; (2) to ensure that minimal effort produces no increase to the effort value; (3) to allow different effort scales.

4.2.1 Document score

For calculating *CE*, an effort score for each ranked document d, $ES(d)$, is needed. The values of $ES(d)$ should increase with the effort; in other words the lower the score the better. There are different possibilities for assigning effort scores for documents. Next, we propose a solution motivated by the evaluation framework.

We assume that the system's task is to point out that the retrieved document is relevant by guiding the user to relevant content. As soon as the user's attention is focused on the relevant spot, the systems mission is accomplished. The document score represents how much expected effort it takes to find relevant text within the document. The scoring depends directly on (non-relevant) characters read before finding the first relevant passage or element. For that we define:

- d' is the expected reading order of document d
- $r_{d'}$ is the position of the first relevant character with the reading order d'

- function $LE(r_{d'})$ gives the localizing effort score based on a chosen window size.

The document effort score $ES(d)$ is the score, that the LE function gives after the relevant text within the document is yielded. For non-relevant documents we assume a default effort score NR:

$$ES(d) = \begin{cases} LE(r_{d'}), & \text{if } d \text{ is relevant} \\ NR, & \text{otherwise} \end{cases}$$

We do not give any default implementation for the LE function. Instead, we introduce sample quantizations in Sect. 5, motivated by the small screen scenario of Sect. 2.

4.2.2 List score

After having defined effort scores for documents in a ranked result list, we can cumulate the effort over the list up to a given cut-off point. Cumulated effort (vector CE) is defined as follows:

$$CE[i] = \sum_{j=1}^{i} \frac{ES(d_j)}{\min ES} - 1$$

where i is a position in the result list and $minES$ denotes the absolute minimum value the function ES delivers. This is obtained when the relevant material is found immediately. The formula ensures that when the effort is minimal the effort value does not increase (cumulate). For instance, let us consider a result list of documents $\langle d_1, d_2, d_3, d_4, d_5 \rangle$ with a vector of corresponding scores $\langle ES(d_1), ES(d_2), ES(d_3), ES(d_4), ES(d_5) \rangle = \langle 1, 2, 5, 1, 5 \rangle$. Moreover, let us assume that the range set of $LE(r_{d'})$ is $\{1, 2, 3, 4\}$ and $NR = 5$, then $minES = 1$. Now $CE = \langle 0, 1, 5, 5, 9 \rangle$.

Normalized cumulated effort (vector NCE) is needed for averaging over multiple topics. It is defined as follows:

$$NCE[i] = \sum_{j=1}^{i} \frac{ES(d_j)}{IE[j]} - 1$$

where IE is the vector representing the ideal performance for the topic. As an example we take the values from the previous example and in addition we state that total number of relevant documents is three, i.e. $Trel = 3$, thus $IE = \langle 1, 1, 1, 5, 5, \ldots \rangle$ and $NCE = \langle 0, 1, 5, 4.2, 4.2, \ldots \rangle$. A normalized optimal run produces a curve having zero values only.

Often it is necessary to have one effectiveness value for the whole result list or a run. An average at a given cut-off point for normalized cumulated effort is calculated as follows:

$$ANCE[i] = \frac{\sum_{j=1}^{i} NCE[j]}{i}$$

where i is the cut-off point. Analogously to mean average precision, mean average normalized cumulated effort ($MANCE[i]$) may be calculated over a set of topics. It is worth noting, that the curves presenting cumulated effort represent the better effectiveness the closer they are to x-axis.

5 Experiments

Next, we illustrate the use of the CE and ChP metrics in testing runs from the RiC task of the INEX 2008 ad hoc track. The RiC task contains 70 topics with character-wise

relevance assessments, and the test collection covers around 660,000 XML marked articles in the English Wikipedia collection (Denoyer and Gallinari 2006). The official results were measured with *F-Score* having alpha value 0.25 (INEX 2009; Kamps et al. 2008c).

Aside from presenting sample results of our metrics and comparing these with the $F_{0.25}$-*Score* metric we aim to study the benefit of using passage retrieval for more effective browsing within the retrieved documents. This is done by comparing the focused fetch and browse strategy with plain full document retrieval. In the document retrieval baseline, the reading starts from the beginning of the document and continues until a relevant passage is met in the *CE* metric, and all relevant passages are read consecutively in *ChPR* metric. This baseline is compared with the corresponding element run.

In Sect. 5.3, based on "Appendix 2", we give a comparative summary of 38 official INEX 2008 runs. First, as a special focus, we report the results of three best performing participants of the RiC task, namely *GPX1CORICe* from the University of Queensland (in Kamps et al. 2008c) and *RICBest* from the University of Waterloo (Itakura and Clarke 2009). For comparison, we selected the best performing full-document run of the task: *manualQEIndri* from the University of Lyon (Ibekwe-SanJuan and SanJuan 2009). Further, we constructed additional runs by transforming GPX1CORICe and RICBest so that the browse phase was discarded, i.e. full documents were returned instead of sets of passages. These runs are labelled as *GPX1CORICe_doc* and *RICBest_doc*.

5.1 Results with character precision-recall

For the Character Precision Recall metric we report results obtained with the following measures: *aveChP* and two T2I based measures, namely $T2If_1(300)$ and $T2If_1(2000)$, where the tolerance of irrelevance is 300 and 2000 characters respectively. In case of the T2I measures the document score is calculated with *F-Score* (note that this is different from *F-Score* of retrieved passages). The α value is 1. In all measures we assume the natural reading order after retrieved passages, i.e. the reading continues from the beginning of the retrieved document after reading the retrieved passages. The *gPr* (list score) curves for the runs GPX1CORICe, RICBest, GPX1CORICe_doc, RICBest_doc and manualQEIndri are shown in Figs. 6, 7 and 8. For comparison, these runs measured with $F_{0.25}$-*Score* of retrieved passages (i.e. the official INEX metric) are shown in Fig. 5. The related *MAgP* values can be found in "Appendix 2".

In Figs. 5, 6, 7 and 8 the superiority of manualQEIndri at early ranks is obvious. It outperforms the element runs, but the comparisons of manualQEIndri with the full document runs (GPX1CORICe_doc and RICBest_doc) show that its better performance is more due to the good document ranking than to full document retrieval competitiveness in focused retrieval (see Kamps et al. 2008a). Adopting focused retrieval clearly gives a boost for RicBest and GPXCORICe in comparison to their full document baselines. This comes especially evident when assuming a lower tolerance to irrelevance, where with the $T2If_1(300)$ measure, the element runs ($MAgP\backslash T2If_1(300)$ 0.187, 0.163, resp.) beat the manualQEIndri (0.151) in addition to their document baselines (0.136, 0.133, resp.). Note that the figures show cut-off results. The differences between document and corresponding element runs by all $MAgP\backslash ChP$ measures are statistically significant ($p < 0.001$, t-test).

5.2 Results with cumulated effort

Measuring the effort on finding relevant content is done with the localizing effort metric for the document score and cumulated effort for the list score. As a basis for calculating the

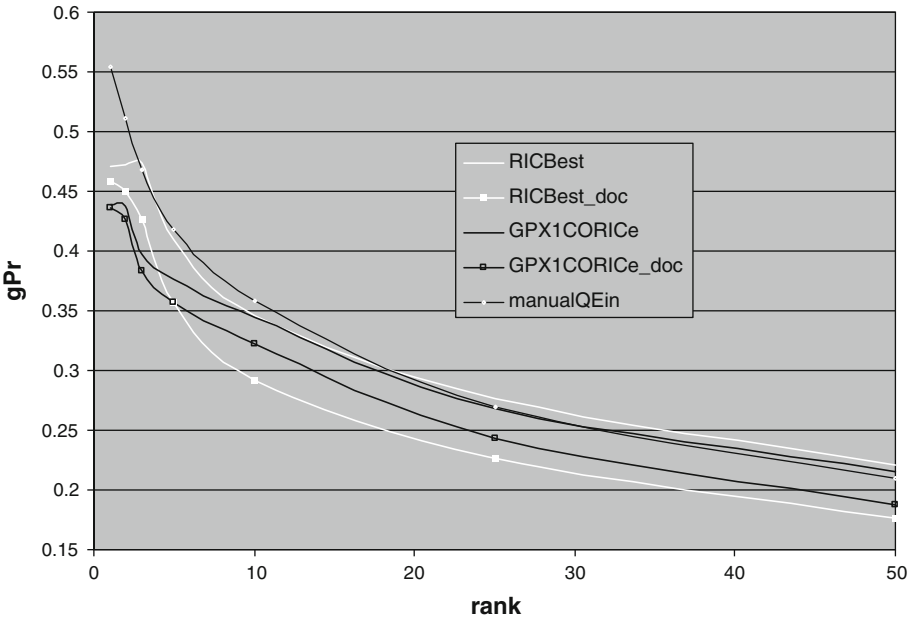

Fig. 5 Generalized precision at cut-off points ($gPr[i]/F_{0.25}$), where the document score is measured with the $F_{0.25}$-*Score*

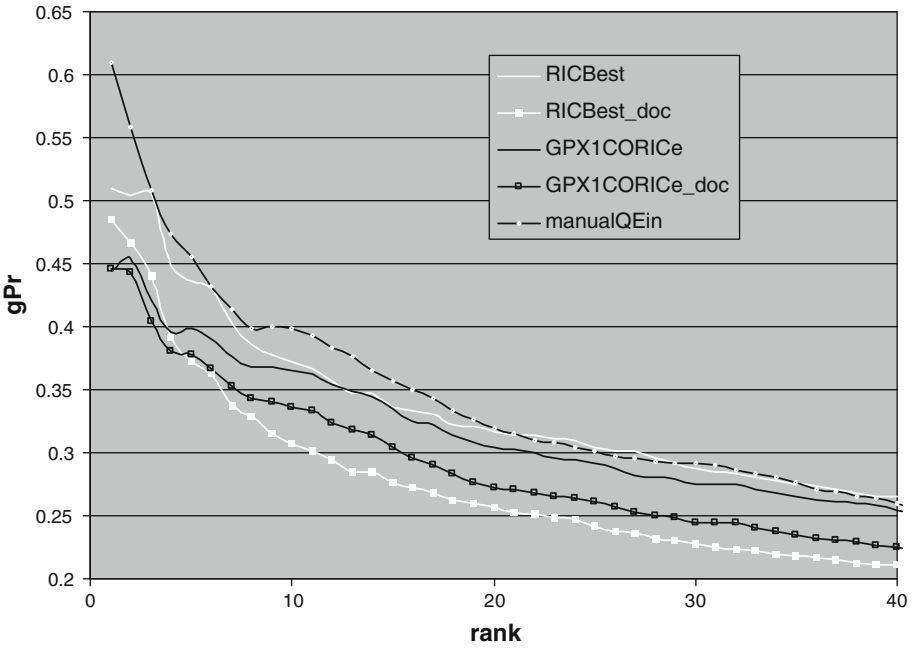

Fig. 6 Generalized precision at cut-off points ($gPr[i]/AveChP$). The document score is measured with the average character precision with the natural case reading order

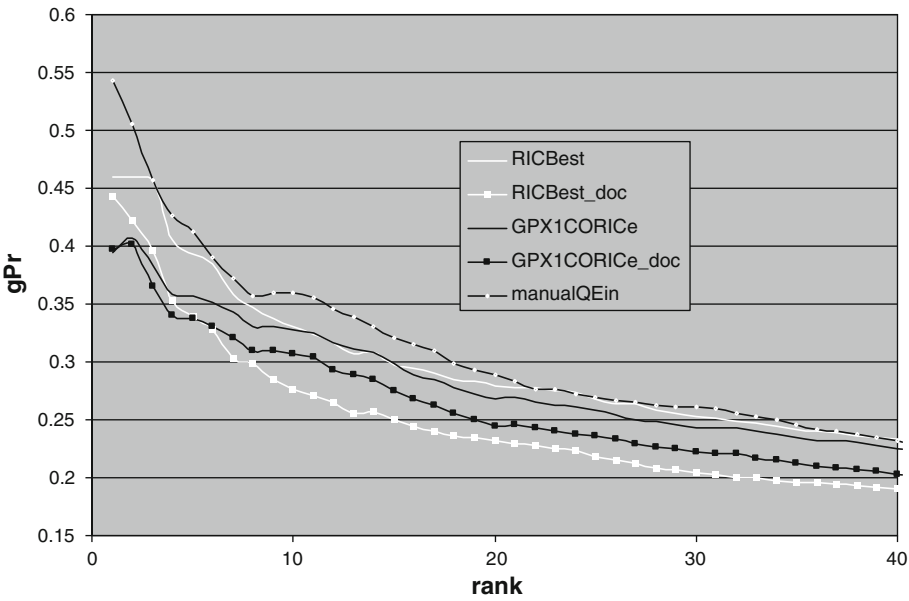

Fig. 7 Generalized precision at cut-off points ($gPr[i]/T2If_1(300)$). The document score is measured with the T2I *F-Score* 300 with the natural case reading order

localizing effort we bind the scoring to the screen size. As scoring for an individual document, we set:

$$LE(i) = \begin{cases} 1, & \text{if } i \leq sSize \\ 2, & \text{if } sSize < i \leq sSize \times 2 \\ 3, & \text{if } sSize \times 2 < i \leq sSize \times 3 \\ 4, & \text{otherwise} \end{cases}$$

$$NR = 5$$

where *sSize* denotes the screen size in characters. For the screen size, we experiment with two distinct values: 300 for a mobile screen and 2000 for a laptop screen. The results are labelled as *screen 300* shown in Fig. 9 and *screen 2000* shown in Fig. 10, respectively. The *MANCE\LE* score of each run is in "Appendix 2". The differences between RicBest and RicBest_doc, as well as GPXCORICe and GPXCORICe_doc are statistically significant ($p < 0.001$, *t*-test) measured with *MANCE\LE*.

The results verify that in comparison to full document retrieval, using a more focused strategy brings somewhat down the effort in localizing the relevant content. Not surprisingly this feature is stressed when using a smaller screen.

5.3 Comparative analysis of the metrics

In addition to comparing top runs, we calculated results for 38 INEX 2008 submissions. In Table 2 Kendall τ correlations of different measures are given. The correlations are based on the results of "Appendix 2". The $F_{0.25}$-*Score* and *ChP* results are calculated with *MAgP* and others with the *MANCE\LE* measure at list cut-off 600. For simplicity the correlations between *MANCE* and *MAgP* are reported as their opposite values, because the score

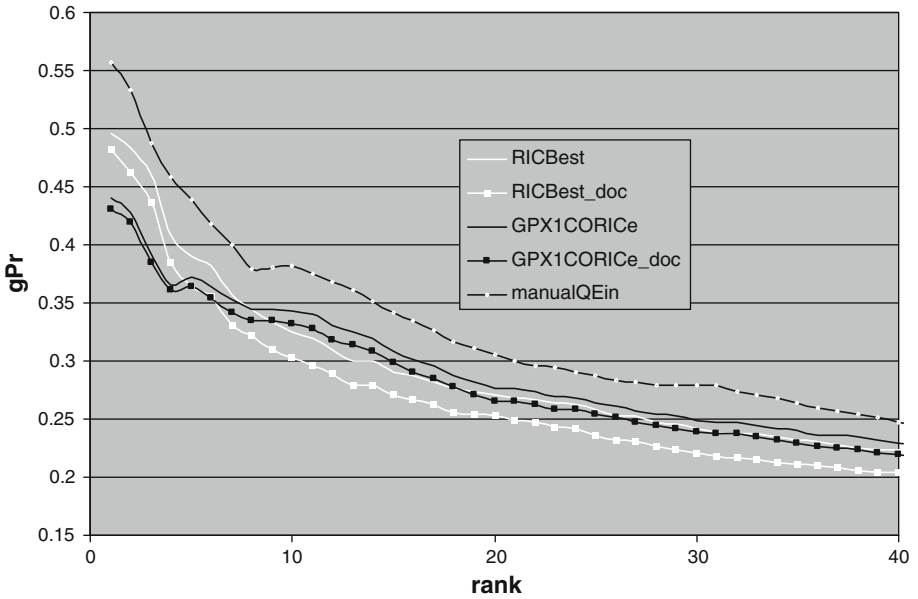

Fig. 8 Generalized precision at cut-off points ($gPr[i]/T2If_1(2000)$). The document score is measured with the T2I *F-Score* 2000 with the natural case reading order

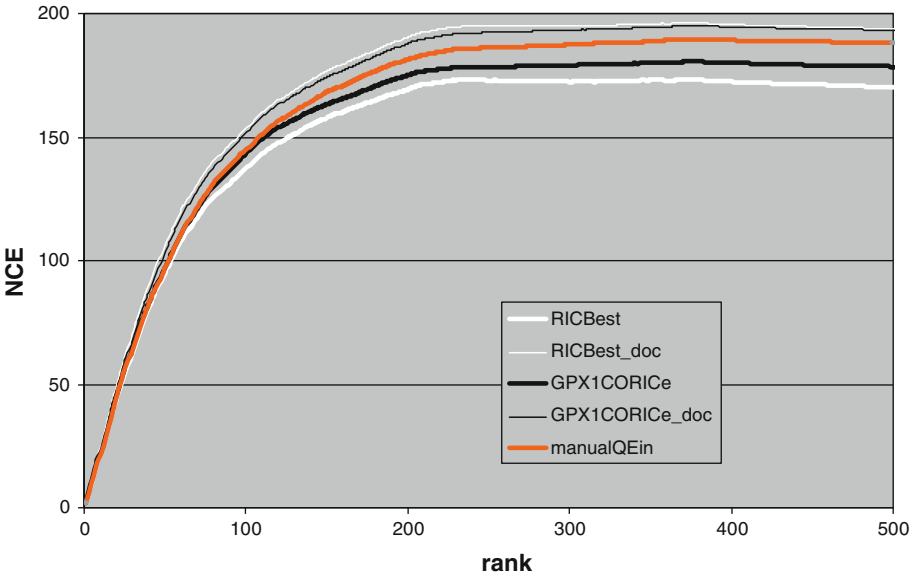

Fig. 9 Normalized cumulated effort with small screen interpretation (screen 300). NB. The lower the curve the less effort spent

interpretations are inverse. In the tables the *doc* ending refers to the document retrieval baseline. For example $F_{0.25}$-*Score doc* means that the runs are handled as if they were full document runs instead of element/passage runs. The correlation between a measure and its counterpart to full document evaluation is in bold.

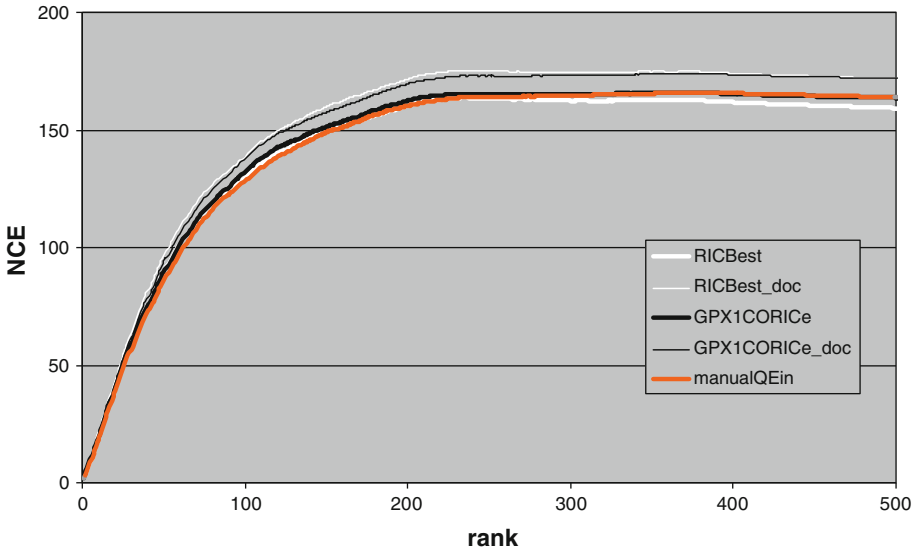

Fig. 10 Normalized cumulated effort with large screen interpretation (screen 2000) NB. The lower the curve the less effort spent

When comparing element/passage runs with their full document baseline, 19 out of 25 runs gain some improvement measured with the $AgP\backslash F_{0.25}$. With Cumulated Effort (for both screen sizes) all runs benefit from the more focused fetch and browse strategy. With *AgP\aveChP* and the reported *AgP\T2I* measures the numbers of benefiting runs are 21 and 15, respectively. The competitiveness of a full document run varies from measure to measure. For instance, the best performing such run, manualQEIndri, is third measured with $MAgP\backslash F_{0.25}$, ninth with the screen size 300 (*MANCE\LE*) and second with the screen size 2000 (*MANCE\LE*). With the *ChP* metric the *MAgP\aveChP* measure delivers third place for the run and with $MAgP\backslash T2If_1(300)$ the ranking is as low as tenth. However, while $T2If_1(300)$ is a rather early cut-off measure (at document level) it might be less reliable, as are the early cut-off measures in general in traditional document retrieval.

6 Discussion

The fetch and browse approach highlights the best matching passages in their context. The aim of this kind of passage retrieval is to make document browsing more effective. In other words the reading order of the retrieved document changes so that the new order is more convenient for the user in comparison to full document retrieval and sequential reading. Thus, successful passage retrieval reduces user effort in finding the best matching parts of the document. In the presented framework the effort is measured with the amount of text the user is supposed to read.

Quite recently the character level of text has been taken into account in the evaluations in the INEX initiative. However, the related *F-Score* metric is system-oriented, and the performance figures are calculated based on the sets of character positions. The set-oriented mindset does not take the reading order into account, which was one of the initial motivations of the fetch and browse style retrieval.

Table 2 Kendall τ correlation of official INEX 2008 results of 38 runs

	MAgP\ $F_{0.25}$ Foc.	*MAgP*\ $F_{0.25}$ Doc.	*MANCE**LE* Screen = 300 Foc.	*MANCE**LE* Screen = 300 Doc.	*MANCE**LE* Screen = 2000 Foc.	*MANCE**LE* Screen = 2000 Doc.	*MAgP*\ *aveChP* Foc.	*MAgP*\ *aveChP* Doc.	*MAgP*\ *T2If*$_1$(300) Foc.	*MAgP*\ *T2If*$_1$(300) Doc.	*MAgP*\ *T2If*$_1$(2000) Foc.
*MAgP**F0.25* Doc	**0.826**										
*MANCE**LE* Screen = 300 Foc.	0.478	0.378									
*MANCE**LE* Screen = 300 Doc.	0.731	0.762	**0.537**								
*MANCE**LE* Screen = 2000 Foc.	0.740	0.629	0.709	0.737							
*MANCE**LE* Screen = 2000 Doc.	0.754	0.768	0.525	0.897	**0.799**						
*MAgP**aveChP* Foc.	0.894	0.789	0.566	0.739	0.817	0.779					
*MAgP**aveChP* Doc.	0.815	0.966	0.372	0.751	0.623	0.762	**0.789**				
*MAgP**T2If*$_1$(300) Foc.	0.695	0.623	0.720	0.697	0.789	0.702	0.795	0.623			
*MAgP**T2If*$_1$(300) Doc.	0.717	0.851	0.408	0.774	0.585	0.711	0.725	0.857	**0.648**		
*MAgP**T2If*$_1$(2000) Foc.	0.837	0.805	0.560	0.776	0.771	0.770	0.897	0.800	0.811	0.773	
*MAgP**T2If*$_1$(2000) Doc.	0.777	0.928	0.405	0.783	0.611	0.749	0.757	0.916	0.628	0.923	**0.805**

We introduced two metrics based on our framework: Character precision-recall (*ChPR*) is based on traditional precision-recall metric. It takes all read text into account also after the first relevant spot and even after the last retrieved passage, if necessary. Within the metric two measures are introduced. Average character precision (*AveChP*) is considered more system-oriented and rewards systems, which are able to present the whole relevant content early to the user. *T2I* based measures takes the user's *tolerance to irrelevance* into account. These measures are based on the total amount of non-relevant characters the user is willing to read per document. Unsurprisingly, the more tolerance to irrelevance the user has the less benefit XML/passage retrieval systems bring.

The cumulated effort metric (*CE*) is a general purpose list measure in a sense that the document scores can be calculated in different ways. In this study the document level measure is localizing effort (*LE*), which measures the effort the user has to take in order to localize the relevant content. In other words, it measures the effectiveness to assess the document to be relevant.

The fetch and browse retrieval is considered a special case of full document retrieval having a flavour of focused retrieval. Thus, good article ranking tends to deliver good results regardless of the metrics. However, the results with the novel metrics showed that the user effort is overall reduced when using passage or XML retrieval. This is illustrated with the pairwise comparisons of element/passage and the corresponding full document run. Thus, the present study gives a partial answer to the concern aroused within the INEX community that the full document retrieval is a competitive approach in fetch and browse style XML retrieval (Kamps et al. 2008a).

Since the experiments were carried out using the existing runs of INEX, any overfitting strategies for the metrics did not show up. As a remote example of returning only the query words within a document might lead to high early precisions at character level. Obviously, the *CE* metrics would deliver good results with that strategy. Clearly, reading a single word is not enough for a user to assess text passages relevance or even to understand it, but he or she has to read the surroundings as well. Therefore, one credible solution preventing this kind of overfitting to the metric would be to set a minimum effort score (penalty) for reading a retrieved passage in addition to the constant effort score reading a character.

Even though we focused on the evaluation of fetch and browse style retrieval, in future studies we will aim to extend this approach to concern other styles of XML and passage retrieval. For instance, instead of starting from the beginning, the browsing of a document may start from the best entry point provided by the IR system (Finesilver and Reid 2003; Reid et al. 2006). This applies to the Best in Context task of INEX. Further, elements can be retrieved as such, i.e. without context. Thus, a result list having elements or arbitrary passages only, like the focused or thorough tasks of INEX, can also be measured within the presented framework.

7 Conclusions

The study gives a framework for the evaluation of element/passage retrieval systems. Unlike the contemporary approaches, the framework is based on reading order, which depends on the co-operative action of a retrieval system and a guiding user interface. The study was motivated by a small screen scenario, where the text is presented as a single column and the default reading of a document is sequential representing the full document retrieval baseline. As the focused retrieval alternative we used a so called fetch and browse approach where effective access to the best matching passages was provided by hyperlinks, still maintaining the document order. Within the scenario we introduced two metrics:

character precision-recall and cumulated effort. In character precision-recall we made an assumption of the user's tolerance to irrelevance, i.e. the point in which the user moves onto the next document. The document score for cumulated effort is calculated with localizing effort function *LE*. In the evaluations we used *LE* functions based on window size motivated by a small screen scenario. However, *LE* can be replaced with other effort measures. We performed laboratory evaluations within the INEX test bed. The results showed that in comparison to traditional full document retrieval, with our measures, more focused element/passage retrieval shows increase in system performance. This gives a better motivation for a fetch and browse style focused retrieval in comparison to the official $F_{0.25}$-*Score* measure.

Acknowledgments The study was supported by Academy of Finland under grants #115480 and #130482.

Appendix 1

See Table 3.

Table 3 List of symbols used in the study

Symbols related to document scoring	
$F\alpha(d)$ *fα*	*F*-Score
$S(d)$	General document score (of document *d*)
$P(d)$	Document precision
$R(d)$	Document recall
$rel(d)$	The set of relevant character positions
$ret(d)$	The set of retrieved character positions
Contribution of this study	
ChPR	Character precision-recall metric
ChP@600	Character precision at cut-off 600
aveChP	Average character precision
$P(p)$	Character precision at position *p*
$RL(p)$	Binary relevance value function of character position *p*
NRC	Number of relevant characters
LE	Localizing effort
NR	Default value for a non-relevant document
$ES(d)$	Effort score (of document *d*)
minES	Absolute minimum of ES function
T2Isco(*300*)	Score (sco) when 300 non-relevant characters are read. (i.e. Tolerance to irrelevance)
Symbols related to list scoring	
gP	Generalized precision
gR	Generalized recall
AgP	Average generalized precision
Trel	Number of relevant documents
Contribution of this study	
CE	Cumulated effort metrics
NCE	Normalized cumulated effort
ANCE	Average normalized cumulated effort
MANCE	Mean average normalized cumulated effort

Appendix 2

See Table 4.

Table 4 Comparison of whole document and passage/element INEX 2008 runs

Run	*MAgP\F*$_{0.25}$			*MANCE\LE* Screen = 300			MANCE\LE Screen = 2000			*MAgP\aveChP*			*MAgP\T2If*$_1$(300)			*MAgP\T2If*$_1$(2000)		
	Doc.	Foc.	Change %	Doc.	Foc.	Change%	Doc.	Foc.	Change %	Doc.	Foc.	Change %	Doc.	Foc.	Change %	Doc.	Foc.	Change %
p78-RICBest	0.188	0.228	21.3	171	152	−12.8	154	143	−7.3	0.202	0.250	23.6	0.136	0.187	38.1	0.168	0.205	22.0
p78-RICArt	0.216	0.227	5.2	162	153	−5.7	148	144	−2.2	0.231	0.249	7.6	0.172	0.196	14.1	0.206	0.221	7.5
p5-GPX1CORICe	0.187	0.211	12.7	169	158	−7.5	152	145	−4.7	0.200	0.228	13.9	0.133	0.163	22.8	0.167	0.190	13.8
p5-GPX2CORICe	0.183	0.206	12.9	173	162	−7.2	156	149	−4.4	0.196	0.224	14.1	0.130	0.160	23.1	0.163	0.185	14.1
p10-TOPXCOallA	0.168	0.195	15.9	171	157	−8.8	153	145	−5.7	0.181	0.216	19.5	0.122	0.154	26.5	0.151	0.176	16.2
p5-GPX3COSRIC	0.168	0.189	12.1	175	164	−6.2	158	152	−3.9	0.180	0.206	14.4	0.117	0.145	23.4	0.150	0.171	13.6
p6-inex08artB	0.158	0.176	11.2	179	168	−7.1	162	156	−4.1	0.173	0.196	13.6	0.112	0.137	23.3	0.141	0.162	14.2
p6-inex08artB	0.165	0.175	6.5	178	169	−5.3	161	157	−2.9	0.180	0.196	8.9	0.118	0.138	17.1	0.149	0.165	10.9
p6-inex08artB	0.164	0.174	6.1	178	170	−5.0	161	157	−2.7	0.180	0.195	8.4	0.118	0.137	16.3	0.148	0.164	10.5
p72-UMDRic2	0.166	0.172	3.6	179	163	−10.3	165	156	−5.7	0.182	0.197	8.2	0.128	0.150	17.5	0.155	0.172	11.1
p6-inex08artB	0.162	0.170	5.1	177	163	−8.4	160	153	−4.3	0.177	0.195	9.9	0.115	0.142	23.6	0.145	0.164	12.7
p6-inex08artB	0.164	0.170	3.5	178	167	−6.6	161	157	−3.1	0.179	0.193	7.3	0.118	0.141	19.4	0.148	0.164	10.8
p6-inex08artB	0.158	0.167	5.9	178	164	−8.6	161	154	−4.6	0.172	0.190	10.3	0.112	0.139	24.1	0.141	0.160	13.0
p72-UMDRic	0.162	0.167	3.3	181	168	−7.5	166	159	−4.7	0.177	0.184	3.9	0.123	0.134	8.4	0.150	0.161	7.1
p12-p8u3exp51	0.135	0.158	17.4	183	166	−10.4	165	157	−5.2	0.148	0.182	23.3	0.098	0.133	35.7	0.122	0.141	15.2
p12-p8u3exp501	0.135	0.158	17.4	183	166	−10.4	165	157	−5.2	0.148	0.182	23.3	0.098	0.133	35.7	0.122	0.138	13.0
p12-p8u3exp311	0.130	0.152	17.0	184	166	−11.0	167	158	−5.3	0.143	0.178	24.2	0.096	0.132	37.8	0.119	0.138	15.7
p48-LIGMLRIC4O	0.154	0.150	−2.5	175	166	−5.4	163	160	−2.1	0.166	0.169	1.9	0.125	0.137	9.7	0.149	0.157	5.7

Table 4 continued

Run	*MAgP\F*$_{0.25}$			*MANCE\LE* Screen = 300			MANCE\LE Screen = 2000			*MAgP\aveChP*			*MAgP\T2If*$_1$(300)			*MAgP\T2If*$_1$(2000)		
	Doc.	Foc.	Change %	Doc.	Foc.	Change%	Doc.	Foc.	Change %	Doc.	Foc.	Change %	Doc.	Foc.	Change %	Doc.	Foc.	Change %
p4-BEIGBEDERI	0.157	0.149	−5.0	177	153	−15.4	158	149	−6.4	0.171	0.206	20.5	0.119	0.165	39.4	0.146	0.167	14.3
p78-RICSum1	0.127	0.141	10.9	184	168	−9.4	171	161	−6.4	0.138	0.159	15.0	0.095	0.119	25.5	0.113	0.134	18.4
p4-BEIGBEDERO	0.106	0.107	0.8	199	181	−9.7	185	176	−4.6	0.115	0.141	22.9	0.078	0.111	41.1	0.097	0.113	16.1
p48-LIGVSMRIC4	0.102	0.096	−5.7	184	181	−1.7	179	177	−0.8	0.110	0.111	0.6	0.093	0.097	4.1	0.102	0.104	2.8
p16-007RunofUn	0.006	0.005	−19.7	293	285	−2.6	288	281	−2.5	0.007	0.006	−13.3	0.006	0.004	−25.6	0.006	0.006	−5.2
p16-009RunofUn	0.005	0.003	−38.5	295	287	−2.8	292	284	−2.8	0.006	0.005	−15.5	0.004	0.003	−26.4	0.005	0.004	−23.4
p16-008RunofUn	0.004	0.003	−36.5	293	285	−2.8	289	281	−2.9	0.006	0.005	−19.3	0.004	0.003	−27.6	0.005	0.004	−19.0
Full Document Runs																		
p92-manualQEin	0.211			164			144			0.232			0.151			0.188		
p4-WHOLEDOC	0.193			170			154			0.207			0.151			0.187		
p4-WHOLEDOCPA	0.193			170			154			0.207			0.151			0.187		
p5-GPX1CORICp	0.190			169			152			0.204			0.135			0.170		
p5-GPX2CORICp	0.187			173			156			0.200			0.133			0.166		
p92-manualweig	0.184			179			163			0.205			0.133			0.165		
p92-autoindri0	0.171			177			158			0.188			0.120			0.153		
p92-manualindr	0.158			183			169			0.171			0.114			0.143		
p5-Terrier	0.152			187			167			0.165			0.103			0.133		
p56-VSMRIP05	0.150			182			164			0.164			0.109			0.133		
p92-manualweig	0.148			188			175			0.163			0.110			0.135		
p56-VSMRIP04	0.131			188			173			0.142			0.091			0.115		
p56-VSMRIP06	0.122			188			171			0.132			0.084			0.108		

References

Ali, M. S., Consens, M. P., Kazai, G., & Lalmas, M. (2008). Structural relevance: A common basis for the evaluation of structured document retrieval. In *Proceedings of CIKM '08* (pp. 1153–1162).

Allan, J. (2004). Hard track overview in TREC 2004: High accuracy retrieval from documents. In *Proceedings of the 13th text retrieval conference (TREC 2004)*. Nist Special Publication, SP 500-261, 11 pages.

Arvola, P., Junkkari, M., & Kekäläinen, J. (2006). Applying XML retrieval methods for result document navigation in small screen devices. *In Proceedings of MobileHCI workshop for ubiquitous information access* (pp. 6–10).

Buyukkokten, O., Garcia-Molina, H., Paepcke, A., & Winograd, T. (2000). Power browser: Efficient web browsing for PDAs. In *Proceedings of CHI '2000* (pp. 430–437).

Chiaramella, Y., Mulhem, P., & Fourel, F. (1996). A model for multimedia search information retrieval. *Technical report, basic research action FERMI 8134*.

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.

de Vries, A. P., Kazai, G., & Lalmas, M. (2004). Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *Proceedings of RIAO 2004* (pp. 463–473).

Denoyer, L., & Gallinari, P. (2006). The Wikipedia XML corpus. *SIGIR Forum, 40*(1), 64–69.

Dunlop, M. D. (1997). Time, relevance and interaction modelling for information retrieval. In *Proceedings of SIGIR '97* (pp. 206–212).

Finesilver K., & Reid J. (2003). User behaviour in the context of structured documents. In *Proceedings of ECIR 2003*, LNCS 2633 (pp. 104–119).

Hyönä, J., & Nurminen, A.-M. (2006). Do adult readers know how they read? Evidence from eye movement patterns and verbal reports. *British Journal of Psychology, 97*(1), 31–50.

Ibekwe-SanJuan, F., & SanJuan, E. (2009). Use of multiword terms and query expansion for interactive information retrieval. In *Advances in Focused Retrieval*, LNCS 5631 (pp. 54–64).

INEX (Initiative for the Evaluation of XML Retrieval) home pages. (2009). Retrieved January 23, 2009 from http://www.inex.otago.ac.nz.

Itakura, K., & Clarke, C. L. K. (2009). University of Waterloo at INEX 2008: Adhoc, book, and link-the-wiki tracks. In *Advances in Focused Retrieval*, LNCS 5631 (pp. 132–139).

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transaction on Information Systems, 20*(4), 422–446.

Jones, M., Buchanan, G., & Mohd-Nasir, N. (1999). Evaluation of WebTwig—a site outliner for handheld Web access. In *Proceedings of international symposium on handheld and ubiquitous computing*, LNCS 1707 (pp. 343–345).

Kamps, J., Geva, S., Trotman, A., Woodley, A., & Koolen, M. (2008c). Overview of the INEX 2008 ad hoc track. In *INEX 2008 workshop pre-proceedings* (pp. 1–28).

Kamps, J., Koolen, M., & Lalmas, M. (2008a). Locating relevant text within XML documents. In *Proceedings of SIGIR'08* (pp. 847–848).

Kamps, J., Lalmas, M., & Pehcevski, J. (2007). Evaluating relevant in context: Document retrieval with a twist. In *Proceedings SIGIR* '07 (pp. 749–750).

Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., & Robertson, S. (2008b). INEX 2007 evaluation measures. In *INEX 2007*, LNCS 4862 (pp. 24–33).

Kazai, G., & Lalmas, M. (2006). Extended cumulated gain measures for the evaluation of content-oriented XML retrieval. *ACM Transaction on Information Systems, 24*(4), 503–542.

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology, 53*, 1120–1129.

Opera Software ASA, Opera Mini™ for Mobile. (2006). Retrieved January 21, 2009 from http://www.opera.com/mini/demo/.

Piwowarski, P. (2006). EPRUM metrics and INEX 2005. In *Proceedings of INEX 2005*, LNCS 3977 (pp. 30–42).

Piwowarski, B., & Dupret, G. (2006). Evaluation in (XML) information retrieval: Expected precision-recall with user modelling (EPRUM). In *Proceedings of SIGIR'06* (pp. 260–267).

Piwowarski, B., & Lalmas, M. (2004). Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *Proceedings of CIKM '04* (pp. 361–370).

Reid, J., Lalmas, M., Finesilver, K., & Hertzum, M. (2006). Best entry points for structured document retrieval: Parts I & II. *Information Processing and Management, 42*, 74–105.

Robertson, S. (2008). A new interpretation of average precision. In *Proceedings of SIGIR '08* (pp. 689–690).

Saracevic, T. (1996). Relevance reconsidered '96. In *Proceedings of CoLIS* (pp. 201–218).
Tombros, A., Larsen, B., & Malik, S. (2005). Report on the INEX 2004 interactive track. *SIGIR Forum, 39*, 43–49.
Trotman, A., Pharo, N., & Lehtonen, M. (2007). XML IR users and use cases. In *Proceedings of INEX 2006*, LNCS 4518 (pp. 400–412).