Expected Ratio of Relevant Units: A Measure for Structured Information Retrieval

Benjamin Piwowarski
LIP 6, Paris, France
bpiwowar@poleia.lip6.fr

Patrick Gallinari
LIP 6, Paris, France
gallinar@poleia.lip6.fr

ABSTRACT

Since the 60's, evaluation has been a key problem for Information Retrieval (IR) systems and has been extensively discussed in the IR community. New IR paradigms, like Structured Information Retrieval (SIR), make classical evaluation measures inappropriate. A few tentative extensions to these measures have been proposed but are also inadequate. We propose in this paper a new measure which is a generalisation of recall. This measure takes into account the specificity of SIR, when elements to be retrieved are linked by structural relationships. We show an instantiation of this measure on the INEX database and present experiments to show how well it is adapted to SIR evaluation.

1. INTRODUCTION

Information Retrieval systems aim at retrieving documents that are *relevant* to a given user information need. The notion of *relevance* is not only not well defined and ambiguous [13, 9], it is also user specific. The evaluation of IR systems appeared very early as a key problem of IR. Cleverdon experiments on the Cranfield collection [3] were the first experiments that justified the development of entirely automatic IR systems. Evaluation is useful for comparing different systems and is used to justify theoretic and/or pragmatic developments of IR systems.

Many different parameters can be used in order to measure the performance of an IR system like for example time and space taken by the system to answer the query and the user effort to find relevant documents. Swets [14] was the first to clearly define how a metric should be defined in order to provide an objective evaluation of IR systems: a measure should only reflect the ability of the system to discriminate relevant documents from irrelevant ones.

A number of hypotheses are also necessary (even if they are implicit) to develop evaluation measures. We can distinguish two kinds of hypotheses: those which are necessary to the computation of the measure and those which are *priors* on user behaviour. Examples of typical assumptions are the following: (1) the user follows the ordered list of retrieved elements beginning with the first element; (2) a relevant document is still relevant even if the user has already seen the same information in another document higher in the retrieved list. We will make such hypotheses explicit when describing our measure.

There are many different approaches for IR evaluation [15, 1]. The expected search length [4] measures the amount of irrelevant documents a user will consult before finding a certain amount of relevant documents. Some measures are based on the definition of a metric over some predefined statistics [2, 15], some derive from rank correlation [10]. But the most famous measures in IR are recall and precision. *Recall* is defined as the ratio of the number of relevant documents that are retrieved to the total number of relevant documents. *Precision* is the ratio of the number of relevant documents that are retrieved to the total number of retrieved documents.

Raghavan [12] proposed a probabilistic version of recall-precision, which is not inconsistent as standard precision/recall can be, especially when documents are not fully ordered. We will not define more precisely their measure here. Instead, we will detail an extension of precision and recall in the case of a non-binary relevance scale, as it was used to evaluate Structured Information Retrieval systems in the 2002 INEX workshop. This extension was proposed by Kekäläinen and Järvelin [7]. In that case, the set R is defined in a fuzzy way: a document can be more or less relevant. When the document is highly relevant, it will be in the set of the relevant documents with a degree of 1. When the document is not relevant, it will be in this set with a degree of 0. Every value between 0 and 1 will be a measure of the relevance of the document. This scale thus generalises the classic binary scale (relevant/not relevant) that is used in IR. Let us denote $j(d)$ the degree with which the document d belongs to the relevant set of documents for a given query. Then, recall and precision are computed as:

$$\text{recall} = \frac{\sum_{e \in L} j(e)}{\sum_{e \in E} j(e)} \quad (1)$$

$$\text{precision} = \frac{\sum_{e \in L} j(e)}{N} \quad (2)$$

where N is the number of documents in the list, E is the set of documents and L is the set of documents in the list. Those two formulas generalise standard recall-precision: when $j(d)$ takes only the values 0 or 1, they give the same results.

In this paper, we propose a measure to evaluate SIR systems. We will first introduce the new problem of SIR. We will show how standard recall/precision have been extended to evaluate such systems and why this is not well adapted to

SIR evaluation. We will then introduce a new measure which is related to the recall. We will compare our measure and precision/recall extension on stereotypical systems using the corpus provided by INEX[1].

1.1 Evaluation and Structured Information Retrieval

Atomic units are usually documents in classical IR. With the actual growth of structured documents [2], the atomic unit is no more the whole document but any logical element in the document. We will call such an element a *doxel* (for DOCument ELement) in the remainder of this paper. Compared to IR on unstructured collections, Structured Information Retrieval (SIR) should not focus on returning documents but *the smallest doxel that contains the answer to the query.* While that query can be only free text like in standard IR (using the INEX terminology, those are *Content Only* queries, CO in short), a query can also specify both constraints on the structure and on the content (those are called *Content And Structure* queries, CAS in short).

We are interested in the evaluation of systems that answer CAS and CO queries, but we will focus here mainly on CO. We will say that a good answer (the smallest doxel) is SIR-relevant to distinguish this notion from usual relevance.

Our work was greatly influenced by the recent INEX initiative [6]. In this section, we describe briefly how SIR systems were evaluated in INEX 2002, which was the first initiative where a corpus of assessed XML documents was built. We will show why the current evaluation methodology is not well suited for SIR.

Let us first describe the INEX scale used for the user assessments. This scale is neither binary, nor between 0 and 1, but is two-dimensional. The first dimension is related to the extent with which the element is relevant. The relevance does not take into account the non relevant part of the doxel, even if that part is 99% of the doxel. For example, the common ancestor of the whole database *will be considered as highly relevant* even if only a small paragraph is highly relevant. In INEX'02, four levels of relevance were distinguished: the doxel can be *irrelevant* (0) if it does not contain any information about the topic of the request; marginally relevant (1) if it mentions the topic of the request, but only in passing; *fairly relevant* (2) if it contains more information than the topic description, but this information is not exhaustive; highly relevant (3) if it discusses the topic of the request exhaustively.

The second dimension, *coverage,* is specific to structured document evaluation. Document coverage describes how much of the document component is relevant to the request topic. Again, there are four levels: no coverage (N) when the query topic is not a theme of the document component; too large (L) when the topic is only a minor theme of the document component; too small (S) when the topic or an aspect of the topic is the main or only theme of the document component, but the component is too small to act as a meaningful unit of information; finally, exact coverage (E) when the topic is the main theme of the doxel.

$$f_g : J_{\text{INEX}} \mapsto J_{[0,1]}$$
$$j \mapsto \begin{cases} 1 & \text{if } j \in \{3E\} \\ 0.75 & \text{if } j \in \{2E, 3L, 3S\} \\ 0.50 & \text{if } j \in \{1E, 2L, 2S\} \\ 0.25 & \text{if } j \in \{1S, 1L\} \\ 0 & \text{if } j \in \{0N\} \end{cases}$$

$$f_s : J_{\text{INEX}} \mapsto J_{[0,1]}$$
$$j \mapsto \begin{cases} 1 & \text{if } j \in \{3E\} \\ 0 & \text{if } j \notin \{3E\} \end{cases}$$

Table 1: Quantisations are used to convert an assessment from the INEX scale J_{INEX} to a binary or real scale used to compute recall and precision. In INEX, two quantisations were proposed: f_s is a "strict" quantisation, $f_g$is a "generalised quantisation"

The two dimensions are not fully independent: a non relevant element (0) must have no coverage (N). There are only 10 different values in this scale (and not 16). In the remainder of this paper, J_{INEX} denotes this set of 10 values. Each of these values is a digit (relevance) followed by a letter (coverage). Thus, $2E$ means "fairly relevant with exact coverage". Within this scale, the doxels that should be returned by a perfect SIR system will be all the doxels *with an exact coverage*, beginning with those with high relevance: in the case of the INEX scale, SIR-relevant doxels are those that have an exact coverage. Doxels with too small or too big coverage in this scale are considered *not relevant.* The motivation is that exact doxels *are* the doxels a user is searching for, while "too small" doxels are contained in an "exact" doxel and "too big" doxels contain an "exact" doxel.

2. LIMITS OF CURRENT MODELS

The first measure proposed in INEX 2002 was standard recall and precision (*i.e.* using f_s, see table 1). In this case, only doxels with exact coverage and high relevance (INEX scale) are the relevant elements (for the binary scale). A system that does *always returns a near match* will have a recall and a precision of 0. This should be avoided since the task complexity is very high. Moreover, when one is assessing the corpus one can find it difficult to give the exact match to one doxel rather than to a smaller one. For example, the list element in INEX often contains only one paragraph; the textual content of both elements (list and paragraph) is thus the same. It is impossible to make a choice and if we give an exact coverage to both, a SIR system will have to return both elements in order to have a perfect recall.

In order to cope with that problem, Gövert [5] proposed to add some relevance to neighbouring doxels, using f_g to convert an assessment from the INEX assessment scale to a value between 0 and 1. A highly relevant doxel with an exact

[1]Initiative for the Evaluation of XML retrieval, http://qmir.dcs.qmw.ac.uk/INEX/

[2]Where the textual (or multimedia) content of the document is usually organised in a tree

match will have a relevance of 1 in the $[0, 1]$ scale. Some of the doxel neighbours will also have a non null relevance: its ancestors – within the document boundary – will have a relevance of 0.75 (too big); some of its children will have a relevance of 0.25 (too small). Non relevant doxel will have a 0 value for relevance. This choice might seem better than the first one, but is still not adequate:

- For every SIR-relevant doxel, there will be a new set of IR-relevant doxels. To give an example of what it implies, consider a system that returns a doxel and two ancestors: this system will have a recall of 2.25, which is better than a system that returns two highly SIR-relevant doxels.
- A system that returns all the SIR-relevant doxels will not be considered as having retrieved all the relevant information: this system will not have a recall of 1.

Those problems are more connected to relevance assessments for free text queries, where there is no constraint on the structure of the retrieved doxels. Nevertheless, the case of structured queries can also be discussed. We will distinguish two different cases:

- The topic formulation does not have any constraint that forbids a doxel and a sub-doxel (a doxel contained in this doxel like e.g. a paragraph in a section) to be both retrieved like for example the query "find a paragraph or a section that talks about cats". Recall/precision are clearly not adapted to this case;
- The topic formulation does not allow a doxel and its sub-doxel to be both retrieved ("chapters that talk about photography"). In this case, we can use standard (or generalised) recall and precision without having any problem.

Classical measures require the definition of the typical behaviour of a system user. This user consults the list of retrieved doxels one by one, beginning with the first returned doxel and continuing in the returned order. In the next section, we propose a measure based on a specific user behaviour, which takes into account the structure of the documents. In particular, we integrated in our measure the fact that a user might explore the doxels which are near the returned doxel in the structure.

In Web-based IR, classical precision/recall can be problematic. Even if the problem is slightly different, some authors have considered using the structural information (hyperlinks) of the corpus. For instance, Quintana, Kamel and McGeachy [11] proposed a measure that takes into account data on the displayed list of documents, on the user knowledge of the topic and also on the links between the documents. They propose to estimate the mean time that a user will spend before finding a relevant document. We follow somewhat the same approach. The main difference is that we rely upon a probabilistic model which makes our measure sound and easily adaptable to new corpora.

3. A MEASURE FOR SIR

We will suppose an ideal situation where assessments in the INEX 2002 corpus *strictly* follow the definition of SIR-relevance (which is not the case). We will thus make the following assumption that a SIR-relevant doxel can only contain SIR-relevant doxels that are less relevant or have a smaller coverage. This constraint states that the same relevant information is assessed with "exact coverage" only one time.

In this section, we describe our measure, beginning with some general hypotheses and its definition. Then we present the probabilistic events and the assumptions we made on them, and finally we show how to calculate our measure.

3.1 Hypotheses

The definition of a measure is based on an hypothetical user behaviour. Hypotheses used in classical measures are subjective but do reflect a reality. In the SIR framework, we will propose a measure that estimates the number of relevant doxels a user might see. We will now describe how a typical user behaves in the context of SIR retrieval. This behaviour will be defined by three different aspects: the doxel list returned by the SIR system, the structure of the documents and the known relevance of doxels to a query. The following hypotheses are similar to that supposed in classical IR:

Order The user follows the list of doxels, beginning with the first returned. He never discourages himself nor does he jump randomly from one doxel to another;

Absolute relevance A doxel is still relevant even if the user has already seen another doxel that contains the same (or a part of the same) information;

Non-additivity Two non relevant doxels will never be relevant even if they are merged.

The three last hypotheses are specific to our measure

Structure browsing The user eventually consults the structural context (parent, children, siblings) of a returned doxel. This hypothesis is related to the inner structure of documents;

Coverage influence The coverage of a doxel influences the behaviour of the user. If the doxel is "too large", then the user will most probably consult its children. If the doxel is "too small", the user will most probably consult the doxel ancestors;

No hyperlink The user will not use any hyperlink. More precisely, he will not jump to another document. This hypothesis is valid in the INEX corpus but can easily be removed in order to cope with hyperlinked corpora.

The measure we propose is the expectation of the number of relevant doxels a user sees when he consults the list of the k first returned doxels divided by the expectation of the number of relevant doxels a user sees if he explores all the

N	Number of doxels in the list consulted by the user
N_R	Number of SIR-relevant doxels that have been seen by the user
L_e	The doxel e is in the list consulted by the user
S_e	The user has seen the doxel e (either in the list or by browsing from a doxel in the list)
$e' \rightarrow e$	The user sees the doxel e after he consulted the doxel e'

Table 2: Events

doxels of the database. We denote this measure by ERR (for Expected Ratio of Relevant documents):

$$ERR = \frac{\mathbb{E}\left[N_R/N = k\right]}{\mathbb{E}\left[N_R/N = |E|\right]}$$

This measure is computed for one query. The measure ERR is normalised ($ERR \in [0,1]$) as $\mathbb{E}\left[N_R/N = |E|\right]$ represents the maximum number of SIR-relevant doxels a user can see in the whole corpus. The measure can thus be averaged over different queries.

3.2 Events

We now have to compute the expectation $\mathbb{E}\left[N_R/N = k\right]$ with the assumptions on the user behaviour we just made. We will introduce some events that are used to formally model the user behaviour and will make some hypotheses on the (probabilistic) relationships between these events. The three different probabilities we introduce are respectively related to the assessments, to the retrieved doxels and to the document structure. The set of events we use in this paper is summarised in table 2.

Events

Let us denote E the set of doxels, e or e' a doxel from E and q a given query. A doxel e can be more or less relevant with respect to the query. We will denote the probability of SIR-relevance of a given doxel by $P(R_e/q)$. The list returned by the SIR system is only partially ordered so that some rearrangements of the list are possible. Depending on the length N of the list, a doxel is then consulted by the user with a probability $P(L_e/q, N = k)$.

When a user consults a doxel e' from the list, he eventually will use the structure to navigate to another doxel e from the document. As it is difficult to make this process deterministic, we will use $P(e' \rightarrow e/q)$ as the probability that the user goes from e' to e. Note that this probability depends upon the query, this will be illustrated in the next sections.

We will suppose that the IR user sees the doxel e iff:

- e is in the list;
- e' is in the list and the user browses from e' to e

This event is denoted S_e and we can write:

$$L_e \vee (\exists e' \in E, L_{e'} \wedge e' \rightarrow e) \equiv \mathrm{S}_e$$

For simplicity, we will now drop the query q from the formulas, as the measure is computed independently for every new query.

Hypotheses

The following hypotheses are necessary for the computation of the measure. Note that all these assumptions are made knowing the query q and the length of the list N. The first two hypotheses are intuitive. The first hypothesis states that the relevance of a doxel does not depend on the fact the user sees it:

$$P(\mathrm{S}_e \wedge R_e) = P(\mathrm{S}_e)P(R_e) \tag{H1}$$

The second states that the behaviour of a user (going from a doxel in the retrieved list to another doxel, $e \rightarrow e'$) does not depend on the fact that the doxel e is in the list (L_e):

$$P(L_{e'} \wedge e' \rightarrow e) = P(L_{e'})P(e' \rightarrow e) \tag{H2}$$

The third states that the fact that events R or L that are related to different doxels are independent, and that in particular

$$\mathrm{S}_e \wedge L_e \text{ or } \neg(\mathrm{S}_e \wedge L_e) \text{ and } \mathrm{S}_{e'} \wedge L_{e'} \text{ or } \neg(\mathrm{S}_{e'} \wedge L_{e'}) \text{ are independant} \tag{H3}$$

This hypothesis has no intuitive meaning and has been introduced only for allowing the measure computation. Nevertheless, it can be justified by those two statements: the relevance is assigned by the user and thus the probability of SIR-relevance does not depend upon the SIR-relevance of another doxel but on the user assessment (that is denoted by our event q). The second point is that the fact S_e that the user sees a doxel e only depends on the fact that a doxel e' is in the list (which is known when we know the length of the list N which is the case here) and that the user moves from a doxel e' in the list to another doxel e.

The third hypothesis is also a simplification of reality, but is as necessary as the two first. It is related to the probability S_e that the user see a doxel e. The more the user can access this doxel from the retrieved doxels by navigating along the document structure, the more "chanches" he has to see that doxel. As it is not possible to evaluate all the interactions between previously seen doxels and this event, we make the hypothesis that correspond to the "noisy-or". This hypothesis is used to compute the probability of the logical implication $A_1 \vee \cdots \vee A_n \Rightarrow B$ as $1 - P(\neg A_1) \ldots P(\neg A_n)$. We thus state that:

$$\begin{aligned} P(\mathrm{S}_e) &= P\left(\textstyle\bigvee_{e' \in E}(L_{e'} \wedge e' \rightarrow e)/N\right) \\ &= 1 - \textstyle\prod_{e' \in E} P\left(\neg(L_{e'} \wedge e' \rightarrow e)/N\right. \end{aligned} \tag{H4}$$

In this equation, we assumed that the event $e \rightarrow e$ is certain (identity move), that is $\mathrm{P}(e \rightarrow e) = 1$ as the logical or is over all doxels in E.

3.3 Theory

In this subsection, we describe how to compute the measure. We now have to derive this measure from the behaviour of a typical user. We will thus introduce a set of probabilities,

each of which describes a part of the user behaviour. We will also make several hypotheses in order to make this measure computable. We now describe several hypotheses that are related to the relevance assessments, to the returned list and to the structure of the documents

We want to calculate $\mathbb{E}[N_R/N = k]$, with $1 \leq k \leq |E|$. We know that by definition,

$$\mathbb{E}[N_R/N = k] = \sum_{r=1}^{|E|} rP(N_R = r/N = k)$$

The user has seen r SIR-relevant doxels ($N_R = r$) when these two conditions are both met: (1) there exists a subset $\{e_1, \ldots, e_r\} \subseteq E$ of SIR relevant doxels that the user has seen and (2) for every other doxel, either the doxel is not SIR-relevant or the user has not seen it. If one considers the set of all sets A that contains r doxels from E, this condition can be written formally as:

$$N_R = r \equiv \bigvee_{\substack{A \subseteq E \\ |A|=r}} \left(\bigwedge_{e \in A} \mathrm{S}_e \wedge R_e \right) \wedge \left(\bigwedge_{e \in E \setminus A} \neg(\mathrm{S}_e \wedge R_e) \right)$$

Events for two different sets are exclusive and using hypothesis (H3) we can state that:

$$\begin{aligned}&\mathbb{E}[N_R/N = k] \\ &= \sum_{r=1}^{|E|} r \sum_{\substack{A \subseteq E \\ |A|=r}} \prod_{e \in A} P(\mathrm{S}_e \wedge R_e/N = k) \\ &\qquad \prod_{e \in E \setminus A} P(\neg(\mathrm{S}_e \wedge R_e)/N = k)\end{aligned}$$

This formula can be reduced, using the hypothesis H1 we obtain:

$$\begin{aligned}\mathbb{E}[N_R/N = k] &= \sum_{e \in E} P(\mathrm{S}_e \wedge R_e/N = k) \\ &= \sum_{e \in E} P(R_e) P(\mathrm{S}_e/N = k)\end{aligned}$$

Using the definition of S_e and the noisy-or hypothesis, we have

$$P(\mathrm{S}_e/N = k) = 1 - \prod_{e' \in E} P\left(\neg(L_{e'} \wedge e' \rightarrow e)/N = k\right)$$

Note that $\mathbb{E}[N_R/N = |E|]$ can easily be computed as $P(S_e/N = |E|) = 1$. Then, using hypothesis (H2), we finally obtain $ERR(k)$:

$$\frac{\sum_{e \in E} P(R_e) \left[1 - \prod_{e' \in E} (1 - P(L_{e'}/N = k) P(e' \rightarrow e)) \right]}{\sum_{e \in E} P(R_e)}$$

3.4 INEX

In the last section, we derived the computation of the measure ERR, but we did not instantiate it in a practical case. We now propose a way to compute some of the probabilities for the INEX database[3], namely for a query the probability $P(R_e)$ of relevance of a doxel and the probability $P(e \rightarrow e')$ that the user browse from a doxel to another.

Computing $P(R_e)$

INEX relevance assessments are given in a two dimensional scale (coverage and relevance). For a given query, we will compute $P(R_e)$ as[4]:

$$P_0(R_e) = \begin{cases} 1 & \text{if } j(e) = 3E \\ 0.5 & \text{if } j(e) = 2E \\ 0.25 & \text{if } j(e) = 1E \\ 0 & \text{otherwise} \end{cases}$$

where $j(e)$ is the assessment of the doxel e for the given query in the scale J_{INEX}. To avoid counting the same relevant information twice, we will furthermore suppose that the probability of SIR-relevance of a doxel is zero whenever the doxel has an ancestor that is relevant with exact match, that is

$$P(R_e) = \begin{cases} 0 & \text{if } \exists e', j(e') \in \{1E, 2E, 3E\} \\ & \text{and } e' \text{ is an ancestor of } e \\ P_0(R_e) & \text{otherwise} \end{cases}$$

Computing $P(e' \rightarrow e)$

To compute the probability that the user jumps from a doxel to another, we will distinguish several relationships between those doxels. Formulas below were only justified by our intuition and can easily be replaced by others. We will denote length(e) the length of doxel e. This length will usually be the number of words contained in the doxel. We will denote by $d(e, e')$ the distance between two doxels. We used the number of words that are between those two doxels: for example, the distance between the last paragraph of section 1 and the second paragraph in section 2 will be the number of words in the first paragraph of section 2 (plus the number of words of the section title). We can now give the formulas, distinguishing four different cases.

e' *and* e *are not in the same document*

We made the hypothesis that the user does not follow any hyperlink:

$$P(e' \rightarrow e) = 0$$

e' *is a descendant of* e

We will suppose that the more e' is an important part of e the greater the probability that a user goes from e' to e. e' relevance has an influence on this probability: if the e'coverage is S (or better, E), the probability is higher:

$$P(e' \rightarrow e) = \left(\frac{|e'|}{\|e\|} \right)^a$$

where a is $\frac{7}{8}$ when the coverage is exact, $\frac{3}{4}$ when the coverage is too small and $\frac{1}{2}$ otherwise.

[3]Note one can use the same definitions for any corpus of structured documents.

[4]Other functions are of course possible, we chose one that seemed "reasonable" to us

e is in e'
This is a symmetric case. The only difference is the coverage influence: a is $\frac{7}{8}$ when the coverage is exact, $\frac{3}{4}$ when the coverage is too big and $\frac{1}{2}$ otherwise.

Other cases
If in the same document two doxels are one after another (like two sibling paragraphs), we will state that the probability that the user follows the path between the two doxel is proportional to the inverse of the distance between the two doxels:

$$P(e' \rightarrow e) = \big(2 + d(e', e)\big)^{-1}$$

4. EXPERIMENTS

4.1 Settings

In this section, we show how the measure discriminates between different IR systems. In order to compare the behaviour of generalised precision-recall versus our measure, we considered six different hypothetical "SIR-systems" which make use of known assessments. These systems exhibit "extreme" behaviours which illustrate a whole set of different situations. The six systems are named:

perfect A system that returns the SIR-relevant doxels

document A system that returns all document in which a SIR-relevant doxel appears

parent A system that returns always the parent of a SIR-relevant document

ancestors A system that returns ancestors of a SIR-relevant document with a score

biggest child The SIR system returns the biggest child (in number of words)

In all these experiments, the score of the doxel was given by the relevance (first dimension of J_{INEX}) of its SIR-relevant doxel: we scored 1 for a doxel which was highly relevant, 0.5 for a fairly relevant doxel and finally 0 for a marginally relevant doxel.

In our experiments, we used all the content only queries for which there were some assessments. We only kept the 1000 first documents returned by the different systems. Given that scores can only take three values, the P/R curve was computed using the Raghavan [12] probabilistic definition of precision and recall (with a step of 0.1). We computed the values at $N = 0...1000$ for our own measure. We averaged our results for P/R and ERR in order to hide the specificities of each assessment. We didn't consider the case of standard precision/recall (e.g. using f_s) as almost all of the models proposed here will have a near null precision-recall curve.

4.2 Results

In figure 1, we present the curves obtained with our measure and in figure 2 the generalised recall/precision (GRP). We will comment on those curves in this subsection: we will point the shortcomings of the GRP and see how our measure corrects the problem. When we analyse those curves, we can at least identify four problems with the GRP:

1. The model **perfect** is not perfect for GRP. This can be seen as it is not the best model and as precision falls very quickly between recall 0.2 and 0.6. This is because when using the generalised quantisation f_g we are adding relevant doxels (for precision/recall) that are not SIR-relevant. Thus, even if the system returns all the SIR-relevant systems, it does not return the other relevant doxels. For our measure ERR, we can see that after almost 400 doxels, model **perfect** has retrieved all SIR-relevant doxels.

2. The model **ancestors** has a higher performance than model **perfect**. This point is related to the previous one: because the model **ancestors** returns more doxels that are relevant (due to the quantisation), recall is better. Due to the limited size of the list and to the 4 possible values for scores, examination of the retrieved doxels shows another thing: every SIR-relevant doxel in the returned list is preceded by a list of its ancestors. We can see this effect with our measure, as the measure increases slowly with the number of the retrieved documents for the model **ancestors**. Our measures also correctly discriminates those two models, as the performance of model **ancestors** is far below the performance of model **perfect**.

3. The model **parent** is much higher than the model **biggest child**. This is not what could be expected, as the parent can contain many irrelevant parts. This effect is due to the fact that doxels with coverage "too small" have a lower value in the real scale than those with coverage "too big". With our measure, model performances are much closer.

4. The model **document** is close to the model **biggest child**. This is not a good property of GRP, since we want a measure that favours systems that retrieved elements of smaller granularity than documents and since the biggest child is very often close to the SIR-relevant doxel (maybe as close as the document). With ERR, this is not the case.

Those four observations show that our measure is better suited to SIR evaluation than GRP. If we consider the theoretic foundations of our measure, it gives some guarantees about its validity.

5. DISCUSSION

In this article, we have described a new measure for SIR systems called the Expected Ratio of Relevant document (ERR). This measure is a generalisation of recall in classical IR: when the probability of going from a doxel to another is always null, the measure reduces to a form of generalised recall. This measure is consistent with SIR, in the sense that it favours systems that find the smallest relevant doxels. Other proposed measures like standard or generalised precision and recall are not good indicators of the performance of a SIR system, as was shown in the last section. Note that results presented here should however be interpreted with care, as we took very specific systems to underline the strange behaviour of GRP. Our measure has the advantage of a sound theoretical foundation and explicitly

integrates the structure of the documents in the modelling of user behaviour[5].

The presented measure could also be very easily adapted in order to evaluate performance of systems in the case of web retrieval. Another interesting property is that it could favour systems that provide Best Entry Points to the document structure [8], from which users can browse to access relevant information. In this case, if from a retrieved doxel there is a high probability that the user goes to some (SIR-)relevant doxels, the measure will increase faster than if the doxel is (SIR-)relevant but provides no (structural) links to other (SIR-)relevant doxels.

The last step would have been to provide an extension of precision as we did for recall. But when we tried to follow the probabilistic approach of Raghavan, a number of problems arose[6] and it is still not clear which set of hypotheses could be used to solve the problem. However, the curves we can draw with the proposed measure are informative enough and have good properties, so it could replace or complement the GRP used for the evaluation of SIR-systems.

[5]This behaviour should be empirically validated.

[6]In particular, we need to calculate the probability of finding N_R relevant doxels in the retrieved list if this list has a given length. This probability can only be computed in $O(2^{MR})$where MR is the number of relevant doxels for the query.

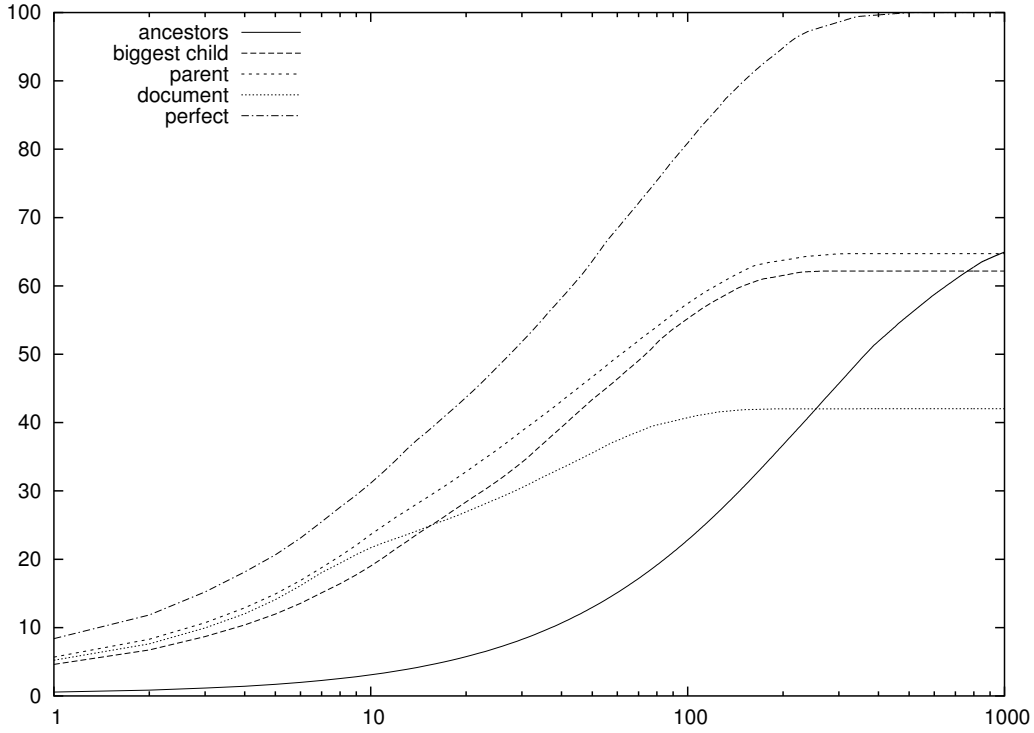

Figure 1: Measure ERR (log-scale for the axis of abscissas). The axis of abscissas represents the length of the list of retrieved doxels. The axis of ordinate represents the measure ERR (in %). The measures are averaged over the queries.

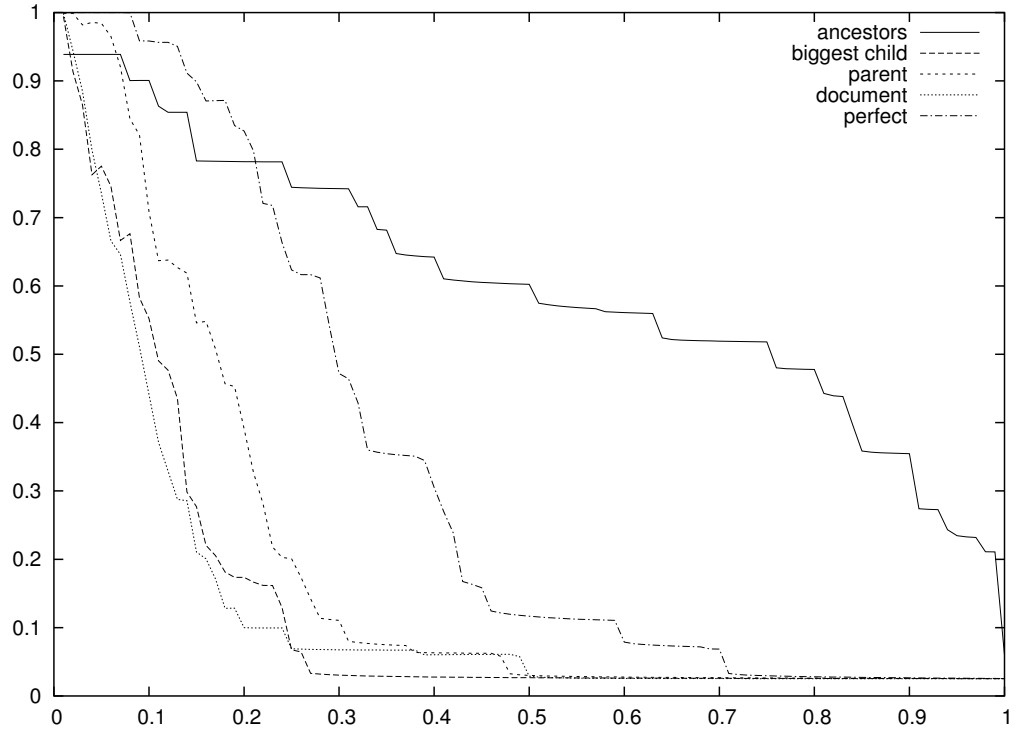

Figure 2: Generalised precision-recall. The axis of abscissas represents recall and the axis of ordinate the precision. Precision are averaged over the queries.

6. REFERENCES

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, New York, USA, 1999.

[2] Peter Bollmann and Vladimir S. Cherniavsky. Measurement-Theoretical Investigation of the MZ-Metric. In Robert N. Oddy, Stephen E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, *Proc. Joint ACM/BCS Symposium in Information Storage and Retrieval*, pages 256–267, 1980.

[3] C.W. Cleverdon. The Cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–192, 1967.

[4] William S. Cooper. Some inconsistencies and misidentified modelling assumptions in probabilistic information retrieval. In Nicholas J. Belkin, Peter Ingwersen, and Annelise Mark Pej, editors, *Proceedings of the 14th ACM SIGIR*, Copenhagen, Danemark, 1992. ACM Press.

[5] Norbert Gövert. Assessments and evaluation measures for XML document retrieval. In *Proceedings of the First Annual Workshop of the Initiative for the Evaluation of XML retrieval (INEX)*, DELOS workshop, Dagstuhl, Germany, December 2002. ERCIM.

[6] Norbert Gövert and Gabriella Kazai. Overview of the Initiative for the Evaluation of XML retrieval (INEX) 2002. In *Proceedings of the First Annual Workshop of the Initiative for the Evaluation of XML retrieval (INEX)*, DELOS workshop, Dagstuhl, Germany, December 2002. ERCIM.

[7] Jaana Kekäläinen and Kalervo Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science (JASIS)*, 53(13):1120–1129, 2002.

[8] Mounia Lalmas and Ekaterini Moutogianni. A Dempster-Shafer indexing for the focussed retrieval of a hierarchically structured document space: Implementation and experiments on a web museum collection. In *6th RIAO Conference, Content-Based Multimedia Information Access*, Paris, France, April 2000.

[9] Stefano Mizzaro. How many relevances in information retrieval? *Interacting With Computers*, 10(3):305–322, 1998.

[10] Stephen M. Pollock. Measures for the Comparison of Information Retrieval Systems. *American Documentation*, 19(3):387–397, October 1968.

[11] Yuri Quintana, Mohamed Kamel, and Rob McGeachy. Formal methods for evaluating information retrieval in hypertext systems. In *Proceedings of the 11th annual international conference on Systems documentation*, pages 259–272, Kitchener-Waterloo, Ontario, Canada, October 1993. ACM Press.

[12] Vijay V. Raghavan, Gwang S. Jung, and Peter Bollmann. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, 7(3):205–229, 1989.

[13] Don R. Swanson. *Historical Note: Information Retrieval and the Future of an Illusion*, pages 555–561. Multimedia Information and Systems. Morgan Kaufmann, July 1997.

[14] John A. Swets. Effectiveness of Information Retrieval Methods. *American Documentation*, 20(1):72–89, January 1969.

[15] Cornelis J. Van Rijsbergen. *Information Retrieval*. Butterworths, 1979.