# Monitoring User-System Performance in Interactive Retrieval Tasks

**Liudmila Boldareva**[†] **Arjen P. de Vries**[‡]
**Djoerd Hiemstra**[†]

[†]University of Twente
Dept. of Computer Science
PO Box 217
7500 AE Enschede
The Netherlands
{l.boldareva,d.hiemstra}@utwente.nl

[‡]CWI
INS1
PO Box 94079
1090 GB Amsterdam
The Netherlands
arjen@cwi.nl

**Abstract**

Monitoring user-system performance in interactive search is a challenging task. Traditional measures of retrieval evaluation, based on recall and precision, are not of any use in real time, for they require a priori knowledge of relevant documents. This paper shows how a Shannon entropy-based measure of user-system performance naturally falls in the framework of (interactive) probabilistic information retrieval. The value of entropy of the distribution of probability of relevance associated with the documents in the collection can be used to monitor search progress in live testing, to allow for example the system to select an optimal combination of search strategies. User profiling and tuning parameters of retrieval systems are other important applications.

**Keywords:**i nformation retrieval, performance monitoring, evaluation, probabilistic retrieval

## 1 Introduction

The goal of an Information Retrieval (IR) system is to satisfy the information need of the user. The term "satisfaction" can be interpreted in more than one way. Depending on the nature of the user's request, the satisfying result might be a complete list of documents related to the query, or a single document[1] containing answers to the request. In some cases the fact that the sought data is missing in the collection is a satisfactory result, e.g., when checking for an existing patent.

The user behind the screen is a key figure in information search. No automatic approach could so far beat a human behind the screen (also when the information need is imposed upon the tester as in TREC evaluations, see more details in Section 4.1). The user participation in the retrieval process is important since even the most advanced indexing techniques are not able to capture all semantics the data represents. Therefore, *interaction* between the user and the system remains a crucial component of information retrieval systems.

[1]In this paper we use term "document" in a broad sense. A document can be any information entity, including text, image, video, structured data and their combination.

An interactive retrieval system should make many decisions during a search session: what objects to show to the user in each iteration, what document representation to use, what collection to search or even whether to give up the search because no relevant answers (seem to) exist. In an environment with long-term learning, when past user input is deployed for system enhancement, it is important to sort out the motivated and successful users from those who randomly browse without particular information need, or without success.

Taking these decisions is only possible if the retrieval system has a means to reflect upon *user-system performance*. For this purpose, this paper proposes a measure to *monitor user-system performance*. This measure should provide the basis for reasoning about the *state* of the search process before and after each iteration, and before and after a search session.

Since long, benchmarking information retrieval systems has measured retrieval effectiveness in *recall* and *precision* (Lesk & Salton, 1968), or a function of those. These are good performance indicators because they intuitively relate to the expected level of user satisfaction. Unfortunately however, computing precision and recall requires knowledge of the relevant set. So, it cannot be used for monitoring the retrieval process, for, if this information were available, we would not be searching!

This paper suggests to relate a user-system performance measure to the user's success at achieving (in cooperation with the system) *search convergence*: i.e., directing the search process through consistent feedback to zoom in upon a specific subset of documents in the collection (hopefully coinciding with the relevant set). In a probabilistic retrieval framework a measure based on Shannon entropy is very convenient for monitoring user-system performance during retrieval sessions. Using entropy as a measure for this purpose does not require knowledge of the relevant subset of documents in the collection, wchich allows for a *passive measurement* of user-system performance. The approach is applicable for any system based on the probabilistic retrieval framework to information retrieval. We give experimental evidence that it is a useful measure for a number of applications in information retrieval system design. The experiments further demonstrate that, in practice, the performance captured by our measure during interactive search sessions correlates with the performance indicated by a precision-recall based measure — mean average precision.

The paper is further organized as follows. The next section reviews briefly the principles underlying a probabilistic approach to information retrieval, and introduce Shannon's entropy and its applications in statistical modeling. Section 3 shows how (the change in) entropy can be used for monitoring of user-system performance, supported by experiment results. The merits of using entropy as the basis for our performance indicator are discussed in Section 5.

## 2 Background

### 2.1 Probabilistic framework in retrieval

A probabilistic model for retrieval predicts the set of documents relevant to the user's information need, based upon their request, possibly accompanied by *relevance feedback*, and the data representation in the storage (also called *indexing*). By using Bayes' rule the problem is stated as estimating for each document the probability of relevance $P(T)$, given the query $Q$, user feedback $F$ and data organisation $I$ (Robertson, 1977; Robertson et al., 1981; Cox et al., 2000; Vasconcelos & Lippman, 2000). We write it down in the following iterative form:

$$P(T) \simeq P(T|Q,F,I) = \frac{P(T)P(Q,F,I|T)}{P(Q,F,I)} \tag{1}$$

where under certain assumptions, the second term in the numerator, $P(Q, F, I|T)$, can be further decomposed into (easier to compute) conditional expressions.

Many approaches exist to estimate $P(T)$; but details of specific retrieval models and their parameter estimation are not important for our purpose here. What is crucial however, is the fact that in a probabilistic framework, the returned documents are ranked according to their (conditional) probability of relevance. The shape of this probability distribution depends on the following conditions, corresponding respectively to random variables $Q$, $F$, $I$ and $T$ (from Eq. (1)):

1. The request posted by the user. With some certainty this request represents the user's information need. We do not know how good this representation is, but we assume that the user is reasonable in his or her query formulation and feedback.
   In an interactive setting, user input indicates relevance of documents augmenting the original request with feedback.

2. The current document representation in the storage. This is often a static component of the model.

3. Prior information about relevance of documents in the collection regardless the actual user request.

### 2.2 Entropy: a reminder

The notion of *entropy* as a measure of uncertainty in a system comes from Information theory. Shannon (1948) defined the entropy in a system with $n$ possible states, with associated probabilities $p_i$, as follows:

$$H = -\sum_{i=1}^{n} p_i \log(p_i) \tag{2}$$

In the absence of constraints and input signals, a uniform distribution for $p_i$ yields the largest value for $H$. When partial information is introduced into the system (e.g., in the IR case, such partial information could be user's query or relevance feedback), the knowledge of the most probable state of the system changes. The *Principle of Maximum Entropy* (`MaxEnt`) predicts that in the new state of the system the entropy takes again the largest achievable value. So, when following the `MaxEnt` principle, the parameters of the distribution of $p_i$ are estimated by finding a (constrained) extremum through maximising Eq. (2). The idea originates from the observation that a mechanical system tends to take a state such that the probability distribution of its phase space has maximal entropy. It has been used for the estimation of (probabilistic) model parameters in many different domains beyond its original application in physics, including speech recognition and natural language processing.

Previous work in IR research applied the `MaxEnt` principle for the estimation of (conditional) probability of relevance that yields the document ranking in the probabilistic framework (Cooper, 1983; Kantor, 1984). Conditions (1)–(3) given in the previous section, specify the set of constraints for solving the maximisation problem for the Information Retrieval domain. These constraints were first identified by Kantor (1984), and later generalised by Greiff and Ponte (2000). Greiff and Ponte have proved that both the Binary Independence Model by (Robertson & Sparck-Jones, 1977) and the Combination Match Model (Croft & Harper, 1979) can be derived from the `MaxEnt` principle.

## 3 Entropy in Interactive Search

The aim of this paper is to demonstrate that the entropy of a probabilistic system has a natural role as indicator of user-system performance. During interactive retrieval the system learns from a user based on his or her feedback. In a probabilistic retrieval system, processing the user feedback leads to the increase of probability of relevance of those documents that might interest the user. Consequently, the goal of a retrieval system is to reach, or to *converge to*, such a state that the probability mass is concentrated on (few) relevant documents, leaving an essentially small (or zero) amount of the mass spread over the remaining non-relevant documents. Assuming that the document collection contains correct answers to a user query, search convergence can be expected to match the user identification of relevant items; otherwise, the user would give feedback that diverges from the current state. Conversely, when several users experiment with the same retrieval system and search task, the entropy in the final state of the search session can be associated with the ability of the user to give consistent relevance feedback.

The argument for suitability of entropy as performance indicator can now be given as follows. Assuming consistent user feedback, search is not very likely to converge *unless* the results satisfy the user's information need. When convergence is reached, the retrieved result set (ranked in accordance with the probability distribution) is thus likely to contain mostly relevant documents. Simultaneously, when search converges, entropy has decreased to a small value: $H$ equals 0 when all documents but one have zero probability of relevance, i.e., the system becomes fully decided.

Summarising the previous discussion, we hypothesise that the entropy $H$ of a probabilistic retrieval system correlates with the quality of:

- the system performance,
- the user who is behind the system,
- and the current progress in retrieval.

The absolute value of $H$ is however only of limited use. First, it is determined (among other factors) by the number of elements in the collection. Second, it is possible to get sporadic low entropy in the data, not related to (desired) peaks in the distribution (see e.g., line with dots in Fig. 1).

For our goal of monitoring user-system performance it seems sufficient to observe how entropy of the system changes. We therefore propose to measure the *change in entropy* at $i^{\text{th}}$ iteration, denoted $\Delta_H^i$.

$$\Delta_H^i = H_i - H_{i-1}.$$

Like $H$ itself, $\Delta_H^{\cdot}$ is easy to track in the ranking process of a probabilistic retrieval system.

Observing $\Delta_H^{\cdot}$ during interactive search allows to draw conclusions about the state of the process. When entropy decreases *consistently* at each iteration, we expect that the user's interaction with the system results in sharpening peaks in the distribution of $P(T)$. The larger the (consistent) decrease in the value of $H$, the better the system distinguishes between (partially) relevant documents and the rest. A descrease that is not supported by further iterations however should be considered coincidental, and does not indicate meaningful progress in the search.

Note that at the beginning of a retrieval session, due to the new input from the user, the entropy is likely to decrease regardless the quality of the user feedback. In Fig. 1 the curves correspond to two types of feedback given by the user. The solid line with dots corresponds to the feedback created by random number generator. The other line denoted as "consistent user feedback" is an example from semi-automatic experiments, where the "user", as stated, gives feedback consistent with the information need. The initial drop in entropy value for the random case is not maintained by the subsequent iterations.

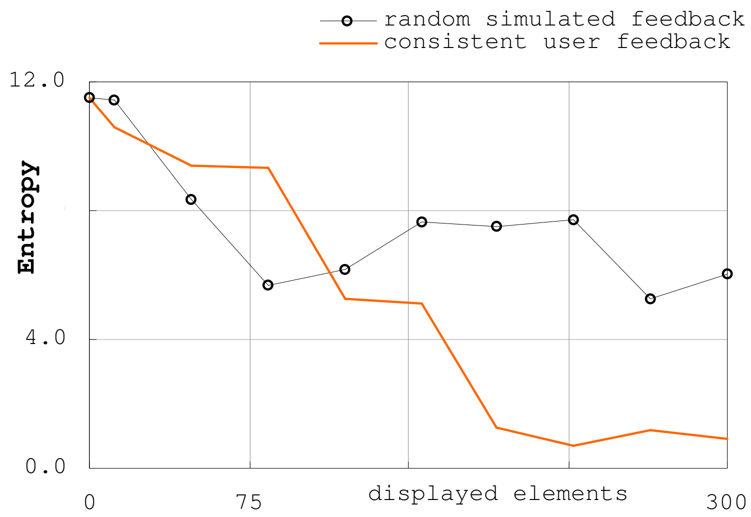


Figure 1: Plots for arbitrary feedback distributed as in real search session vs. consistent feedback from experiments

## 4 Experiments

To support our considerations about the importance of entropy, we analysed experiments performed on TREC Video interactive search task (see also (Voorhees & Buckland, 2003)). An interactive set-up is perfectly suited for search and retrieval of visual data, for a number of reasons. First, indexing of video data is still far from optimal. Capturing content of an image or a video-clip has advanced in its best to a two-step process: labeling a training set by humans, followed by the application of machine learning techniques to propagate to the whole collection. Fully automatic methods for multimedia retrieval are way behind the progress made in the textual domain. Thus, the user input to the search is indispensable. Second, providing a good query or visual example might be a problem on its own, while expressing the visual aspects in the user information need through feedback is quite convenient. Finally, visual data is particularly suited for interaction. Judging the relevance of an image takes a fraction of a second, while assessing a piece of text requires at least reading part of the results; even if only scanning the text for presence of (highlighted) query words.

The aim of the experiments presented here is:

- To show that entropy change gives information about search success;
- To demonstrate the relationship between the change in entropy and convergence of search to a set of (mostly relevant) documents.

### 4.1 Interactive Experiment Setup

We use video data provided by TREC in the framework of TREC Video workshop (TRECVID). The videos are divided into shots, and from each shot a key frame is extracted. Visual similarity between the key-frames is computed using a generative probabilistic retrieval model (see for details (Westerveld et al., 2003)). The probabilistic approach used for indexing and retrieval has been described in (Boldareva et al., 2003). Text from the search topic descriptions as well as speech transcripts of example videos is combined with visual similarities between key-frames to contribute to the probability distribution $P(T)$ used for ranking, along with the relevance feedback. We developed several ways to perform *display update*, i.e., composing new set of images to be presented to the user for relevance feedback. Mean average precision (`MAP`) was used as a measure to user's satisfaction. To calculate average precision for one search session, precision for each retrieved relevant document is calculated, and then the average is taken. For all search tasks of TREC, a mean value is taken to produce a single number – `MAP`.

In the first series of experiments, referred to as *semi-automatic*, user input has been simulated by using relevance judgements available from TREC assessors, making up in this way for a "generic user". In a second series of experiments, the *live* experiments, the TRECVID 2003 search topics have been performed by real users. The testers were free to browse the collection, and relied on their own notion of relevance. We found high agreement between real users' feedback and the TREC relevance judgements (average among users 75%), so our semi-automatic experiments can be considered a good approximation to real life (see (Voorhees, 2000) for an analysis of agreement between TREC assessors). The goal of the semi-automatic setup has been to find the best combination of search strategies in our probabilistic framework.

### 4.2 Semi-automatic experiments

The semi-automatic experiments have been performed on a subset of the TRECVID collection, to achieve a larger proportion of relevant documents (half of the key frames was relevant to at least one of 25 topics, with the average of 2% per topic).

We calculated `MAP` based on the ranking produced by the modelled distribution $P(T)$, every three iterations. Recall that `MAP` serves to model user satisfaction. For the same iterations we computed the value of entropy. Example plots for two search methods (averaged over 25 search queries) can be found in Fig. 2. A similar dependency between the two measures is observed for other set-ups. To make explicit how well the change of entropy predicts `MAP`, we computed the correlation coefficient between the decrease in entropy and increase in mean average precision. The correlation estimate varied between 0.79 and 0.99 for different set-ups, with the overall value of 0.85. This strong correlation between change in entropy and `MAP` provides statistical evidence that the change in entropy serves indeed as a suitable indicator of search progress.

To emulate another situation in which we would like to use the change in entropy for monitoring user-system performance, we performed the following semi-automated experiment: during a session, the search topic was changed, but the feedback related to earlier topics, was not forgotten. The setup was repeated for different search strategies. In addition, we switched between "easy" topics (those that had proportion up to 7–10% of relevant items) and the "hard" ones that counted a meager amount (0.5–0.2%) of relevant examples. The summary of the experiment is plotted in Fig. 3. The change in user's perception is clearly reflected by the dynamics of entropy change. We conclude that it should be possible by tracking the changes in entropy to identify when users change their search strategy, or even, as shown here, change the search target in their mind.

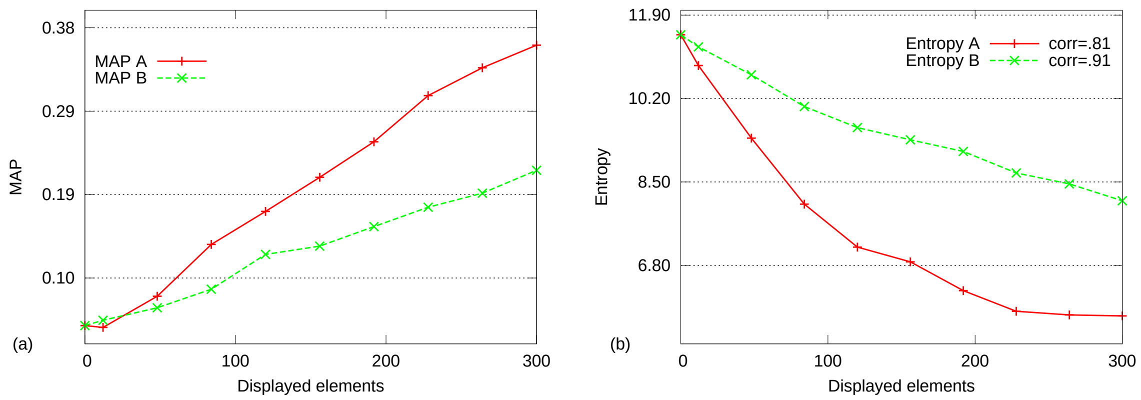


Figure 2: Average precision (a) and corresponding entropy (b) value for two methods, correlation between increments shown in (b).

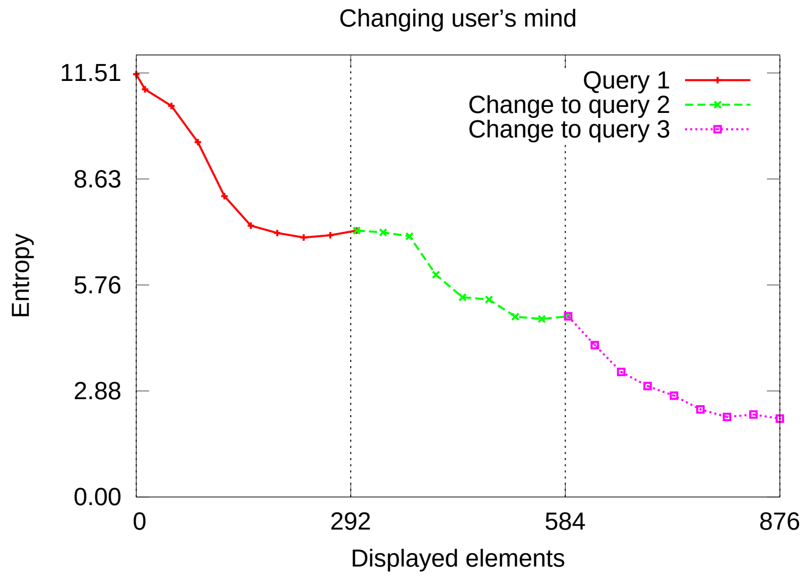


Figure 3: Changing the goal during one search session. Grid-lines correspond to the moments of switching.

### 4.3 Live experiments

For search experiments with real users there were 2 groups of 3 students to test systems A, B and C on 24 search tasks. For a group, each system was tested by three different users, and, in turn, each user had 8 topics from each system to test ("Latin square" experiment design). No user had to perform the same search task twice. The positive feedback from the user was added to the query, and the next screen consisted of nearest neighbours of the positive visual examples.

All users did their best to keep consistent feedback, and the search task description was available at any moment during the search. However, errors and glitches in feedback caused by human factor were unavoidable.

The TREC evaluation guidelines required that each tester performed all search topics. Because we wanted to submit our best results, we attempted to identify from each group those users that were overall most efficient. Following the reasoning explained before, we predicted that our "best" users should have reached the system state with the lowest entropy value at the end of each search session. The number of times that each user turned out as best with respect to the final entropy value was chosen as the criterion for selection whose search results would be part of our submitted interactive run.

The search results of more efficient users (group II) and less efficient users (group I) were then re-combined to make 3 Systems: A, B, or C, in accordance with TREC requirement. For each system a standard evaluation measure (`MAP`) was available later. Table 1 contains numbers for the entropy-based score and corresponding mean average precision for the two groups I and II, and two systems B and C (since system A used randomised search strategy, it was not taken into account when calculating scores). Note that the selection into group II has been made based on the data available at submission time, *which did not include the knowledge of the relevance sets*. In Table 2 the TREC requirement is abandoned, and the search results for evaluation are selected for each topic individually. Analogously, the search outcomes were re-combined into two groups based on per-topic efficiency with respect to the entropy, namely, less efficient group III versus more efficient group IV. The resulting difference between the groups is even more convincing.

| System Group I | Entropy-based score | `MAP` value | | System Group II | Entropy-based score | `MAP` value |
|---|---|---|---|---|---|---|
| B–I | 11 | 0.204 | vs. | B–II | 13 | **0.214** |
| C–I | 7 | 0.243 | vs. | C–II | 17 | **0.245** |

Table 1: Entropy-based scores and performance figures for real users in TREC, grouped by the system types B, C based on overall efficiency of users

| System Group III | Entropy-based score | `MAP` value | | System Group IV | Entropy-based score | `MAP` value |
|---|---|---|---|---|---|---|
| B–III | 0 | 0.199 | vs. | B–IV | 24 | **0.218** |
| C–III | 0 | 0.234 | vs. | C–IV | 24 | **0.255** |

Table 2: Entropy-based scores and performance figures for real users in TREC, grouped by the system type B, C based on per-topic efficiency of users

## 5 Discussion

We showed that our entropy-based method is a useful tool to measure search success. Using experiments on TREC video data, we showed that entropy can be used in three main contexts:

– To evaluate differences of performance when tweaking parameters of retrieval system without changing the model constraints.

– To classify users based on the (in-)consistency in his or her search behaviour.

– To track the user-system progress in real-time. The entropy change bears information about effectiveness of current search strategy.

The results have demonstrated that entropy change strongly correlates with a change in mean average precision. When the increase of entropy tends to zero, this is an indication that the user transfered to the system maximal knowledge about his or her information need, and further improvement of the search results is unlikely.

What we can measure is how well a system reacts to (consistent) user feedback. This allows a system to reflect upon its own operation, and explain to the user whether it can "make sense" out of the user feedback. This is a very desirable property, as such a dialogue between user and system can significantly improve the user's feeling of being in control. And, giving the user sense that they are in charge of the system is one of Shneiderman's "Eight Golden Rules of Interface design" that make a satisfied user (Shneiderman, 1998).

The second result has two applications. First, it allows a posteriori selection of more efficient users in test environment, such that their input can then be introduced into the retrieval model, contributing to "long-term learning from user interaction". Second, it enables "on-the-fly" detection of those cases when the user's information need cannot be satisfied in the collection, or at least not by this combination of user and retrieval system. In this case certain steps can be taken *during* the search to resolve this situation.

The third item motivates further research in the application of this measure for dynamically adjusting the search strategy to the observed user behaviour. In many cases, the best search performance is achieved by a *combination* of search strategies, especially in multimedia search. The monitoring of user-system performance with measure $\dot{\Delta}_H$ allows the system to adapt its behaviour to the situation encountered in real time. Based on change of entropy, it can detect when the retrieval strategy is no longer productive and should be changed: e.g., to move away from a localised search that "got stuck", by starting a fresh (global) session. For instance, when a text query in a video retrieval task does not result in any measurable convergence in the user-system cooperation, it is reasonable to sample the whole data collection, gradually moving to a nearest neighbour strategy using visual similarity, for the refined set of relevant documents. Ultimately the system can point out to the user that the feedback seems inconsistent regarding current data organisation, so there may be no answers to his or her query.

To sum up, our usage of entropy as a performance indicator in interactive IR has the following attractive properties:

- it does not depend upon knowledge of the relevant documents;
- its computation is inexpensive;
- and, therefore, it allows for a timely response to "troubling symptoms" in the interaction.

Measuring the dynamics of entropy provides a key component to turn the interaction between user and system in interactive information retrieval into a true dialogue!

## References


Boldareva, L., Hiemstra, D., & Jonker, W. (2003). Relevance feedback in probabilistic multimedia retrieval. In *DELOS-NSF workshop on multimedia contents in digital libraries.* http://www.music.tuc.gr/MDL/.

Cooper, W. (1983). Exploiting the maximum entropy principle to increase retrieval effectiveness. *Journal of the American Society for Information Science*, *34*(1), 31–39.

Cox, I. J., Miller, M. L., Minka, T. P., Papathomas, T. V., & Yianilos, P. N. (2000). The bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Tran. On Image Processing*, *9*(1), 20-37.

Croft, W., & Harper, D. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, *35*(4), 285–295.

Greiff, W. R., & Ponte, J. M. (2000). The maximum entropy approach and probabilistic IR models. *ACM Trans. on Information Systems*, *18*(3), 246-287.

Kantor, P. B. (1984). Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology, Research and Development*, *2*(3), 88–94.

Lesk, M. E., & Salton, G. (1968). Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, *3*(4), 343–359.

Robertson, S., & Sparck-Jones, K. (1977). Relevance weighting of search terms. *Journal of the American Society for Information Science*, *27*, 129–146.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, *33*(4), 294-304.

Robertson, S. E., van Rijsbergen, C. J., & Porter, M. F. (1981). Probabilistic models of indexing and searching. In *Proceedings of the 3rd annual acm conference on research and development in information retrieval* (pp. 35–56). Butterworth & Co.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379-423, 623-656.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (3rd ed. ed.). Menlo Park, C: Addison Wesley.

Vasconcelos, N. M., & Lippman, A. B. (2000). A probabilistic architecture for content-based image retrieval. In *Proc. of IEEE conf. on computer vision and pattern recognition.*

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing Management*, *36*(5), 697–716.

Voorhees, E. M., & Buckland, L. P. (Eds.). (2003). *The twelveth text retrieval conference TREC-2003,gaithersburg.*

Westerveld, T., Vries, A. de, Ballegooij, A. van, Jong, F. de, & Hiemstra, D. (2003). A probabilistic multimedia retrieval model and its evaluation. *EURASIP Journal on Applied Signal Processing, special issue on Unstructured Information Management from Multimedia Data Sources*, *2003*(2), 186–198.