A Geometric View of Relevance Effectiveness in Information Retrieval

Sándor Dominich

*Department of Computer Science and Information Technology, Buckinghamshire Chilterns University College, High Wycombe, HP*11 *2JZ, Buckinghamshire, UK, E-mail: sdomin*01*@buckscol.ac.uk; Department of Computer Science, University of Veszprem, Egyetem u. 10, 8200 Veszprem, Hungary, E-mail: dominich@dcs.vein.hu*[1]

Abstract. Relevance is a central concept in Information Retrieval (IR). It is used to work out effectiveness measures for IR systems, i.e. measures to express how well (or bad) an IR system performs; classical measures are precision, recall, fallout. It is shown that the empirical relation $P=NR/x$ (P=precision, R=recall, N=total number of relevant documents, x=the number of retrieved documents) can be formally easily obtained. It is also shown that using the concept of fallout a typical surface can be constructed with the noteworthy properties that it looks similarly for every IR system and each point on this surface corresponds to a 3-tuple (precision, recall, fallout) and thus to one retrieval process. Thus, the name of effectiveness surface is suggested for it. The performance of an IR system can be enhanced by a technique called relevance feedback (used to return documents that are likely to be more relevant). A sequence of repeatedly applied relevance feedbacks, being a sequence of repeated retrievals, corresponds to a sequence of points ('walk') on the effectiveness surface. It is shown that this sequence can be theoretically modelled by an important mathematical structure (recursively enumerable set or Diophantine set), and that it yields a point on the effectiveness surface corresponding to an optimal retrieval situation. Further, the existence of an optimal point is also shown and it is computed as well.

Keywords**:** Information Retrieval, Relevance Theory, Constrained Nonlinear Optimisation.

1. Introduction

Relevance is a central concept in Information Retrieval (IR), and is mainly a subjective category expressing the user's judgment as to how relevant (or irrelevant) a returned document is. A technique called *relevance feedback* can be used to return documents that are likely to be more relevant: the user is presented with a list of documents first and asked to rank them according to his/her relevance judgment; this is used then to return other documents that are likely to better satisfy his/her information need. Because of the central role played by relevance, it has been and is used to work out *effectiveness measures* for IR systems, i.e. measures to express how well (or bad) an IR system performs. The complexity of this task as well as the compoudness of relevance is also well reflected in recent years research by e.g. Mizzaro (1997), Allan (1996), Belkin (1996), Belkin & Koenemann (1996), JASIS (1994).

Precision and *recall* are two traditional effectiveness measures: precision means the proportion of relevant documents out of those returned, whereas recall that of returned documents out of the relevant ones. Buckland & Gey (1994) attempt to elaborate polynomials

[1] Research partly supported by research grant Pro Cultura Renovanda Tud 98/31

for precision and recall using postulates based on empirical considerations. Their result is that $P = N_{rel} R / x$, where P means precision, R means recall, N_{rel} is the total number of relevant documents and x is the number of retrieved documents.

In the present paper it is shown that this relationship can be easily obtained in a formal (theoretical) way, too. Thus, this may be used as a theoretical introduction to Buckland and Gey's cited work. Moreover, using a third traditional concept for effectiveness, *fallout* (which means the proportion of returned documents out of those nonrelevant), it is shown that a surface can be constructed with the following properties:

a) It looks similarly for every IR system (only its specific position, not its shape, in space varies depending on the total number of documents and the total number of relevant documents given a query).

b) Each point on this surface corresponds to a 3-tuple (precision, recall, fallout) and thus to one retrieval process.

A sequence of repeatedly applied relevance feedbacks corresponds to a sequence of points on this surface. Assuming that the sequence of relevance feedbacks (e.g. in a probabilistic retrieval) improves the effectiveness of the IR system, i.e. it improves precision, recall and fallout, the question of whether the corresponding sequence of points tends towards a specific point (corresponding to an optimum) arises, and if so which is that point? In other words, the sequence of retrievals means a 'walk' on the surface. In mathematical terms, the above question means the following: is there an optimal (minimal) point that this sequence of points leads to? The answer to this question will be yes: it will be shown that a repeatedly applied sequence of relevance feedbacks in probabilistic retrieval can be conceived as a recursive process that can be theoretically modelled by an important mathematical structure (recursively enumerable set or Diophantine set). A noteworthy property of such a structure is that it has a fixed point. This point can be thought of as corresponding to a retrieval situation which cannot be improved anymore, hence the existence of an optimal point is granted and it can be computed. This computation will also be performed.

The paper suggests that this surface be called an *effectiveness surface* which thus offers a geometric view of relevance effectiveness of an IR system. At the same time, it also reflects a theoretical dynamics among precision, recall and fallout during repeated relevance feedbacks.

2. Definitions of the main mathematical concepts used

1. *Recursion* means to define a process (function, procedure, language) with reference to itself. Formally, recursion is a process whereby a function *f*, called a primitive recursive function, with n+1 variables is defined as follows: $f(x_1, x_2, ..., x_n, 0) = \alpha(x_1, x_2, ..., x_n)$, $f(x_1, x_2, ..., x_n, y+1) = \beta(x_1, x_2, ..., x_n, y, f(x_1, x_2, ..., x_n, y))$ where α is a function with n variables and β a function with n+2 variables. In words, function *f* is given an initial value first represented by function α, and then every next value of *f* is defined by function β using the previous value of function *f*. An example for recursion is the usual arithmetical addition: $f(x, 0) = x$, $f(x, y+1) = f(x, y) + 1$. Let us consider the particular case when n=1. Then $f(x, 0) = \alpha(x)$, $f(x, y+1) = \beta(x, y, f(x, y))$, where $x = x_1$ and hence the index can be omitted. This particular case is used in Theorem 1.

2. *Fixed point*. Given a recursion with a corresponding *f* (as above). Then there exists a situation when the value of *f* coincides with the value on which *f* is computed; symbolically $f(x) = x$. This value is referred to as a fixed point. There are several fixed points theorems in recursion theory (e.g. Rogers Fixed Point, First Recursion and Second Recursion Theorems). For this paper, it is important that a fixed point exists, and this corresponds to a situation where a retrieved set of documents is the same as that retrieved in the next step (after relevance feedback). This may, in principle, be interpreted as a case when effectiveness cannot be enhanced anymore. This concept and interpretation of a fixed point is used in part 4.

3. *Diophantine set* (or recursively enumerable set, abbreviation; r.e. or Diophantine structure). A subset $A \subseteq B$ is called a recursively enumerable or Diphantine set relative to *B* if there exists a procedure (process, algorithm, program, language) that, when presented with an input $b \in B$, outputs 'yes' if and only if $b \in A$. When $b \notin A$ then the procedure does not end (undefined). For example, the set of C programs that halt on a given input is a r.e. set (relative to the set of all C programs). It can be shown that a set whose elements are given (generated) by a primitive recursive function is a Diophantine set. This result is used in Part 4.

4. *Level surface*. Given a function $f: \mathbf{R}^3 \rightarrow \mathbf{R}$, $f(x, y, z) \in \mathbf{R}$, $(x, y, z) \in \mathbf{R}^3$, and a constant $c \in \mathbf{R}$. The set of points (x, y, z) in space for which $f(x, y, z) = c$ is called a *level surface*. For example, let $f(x, y, z) = x^2 + y^2 + z^2$ and $c = 4$. Then the level surface $x^2 + y^2 + z^2 = 4$ is the sphere having its centre in the origin and radius equal to 2. All level surfaces in this example are spheres but with different radii. This result is used in Part 3.

3. Effectiveness surface

Let $\Delta \neq 0$ denote the total number of relevant documents to a query q, $|\Re(q)| = \kappa \neq 0$ denote the number of retrieved documents in response to q, and α denote the number of retrieved and relevant documents. It is reasonable to assume that $|D| = M > \Delta$.

The meaning of the usual relevance effectiveness measures are recalled now as follows.

DEFINITION 1. *Recall* ρ is defined as $\rho = \alpha / \Delta$.
DEFINITION 2. *Precision* π is defined as $\pi = \alpha / \kappa$.

One can easily see that:

PROPOSITION 1. *The ratio of recall and precision varies linearly with* κ.

(Proof. $\alpha = \rho\Delta = \pi\kappa \Rightarrow \ \rho / \pi \ = \ \kappa / \Delta$.)

The result reported by Buckland and Gey (1994) is $P = N_{rel} R / x$. This rewrites as $R / P = x / N_{rel}$ which coincides with that of Proposition 1.

A third traditional measure is as follows:

DEFINITION 3. *Fallout* φ is defined as $\varphi = (\kappa - \alpha) / (M - \Delta)$.

As a noteworthy property, it can be shown now that recall, precision and fallout satisfy the following relation:

PROPOSITION 2.

$$\frac{\varphi\pi}{\rho(1-\pi)} = \frac{\Delta}{M-\Delta}$$

Proof.

$$\varphi = \frac{\kappa - \alpha}{M-\Delta} = \frac{\kappa - \rho\Delta}{M-\Delta} = \frac{\rho\Delta/\pi \ - \ \rho\Delta}{M-\Delta} = \frac{\rho\Delta(1-\pi)}{\pi(M-\Delta)}$$

Because recall ρ, fallout φ and precision π can take on different values e.g. in a series of relevance feedbacks for the same query q, they can be thought of as values of variables in general as follows: x for fallout, y for precision and z for recall. Thus the left hand side of the expression in Proposition 2 corresponds to a three-variables function $f(x, y, z) = xy / (z(1 - y))$. The right hand side of the above expression is constant for a query q under consideration. Thus, one can consider all those values of function f, i.e. points in the three dimensional Euclidean space, that are equal to the right hand side. This is equivalent to defining a surface as follows.

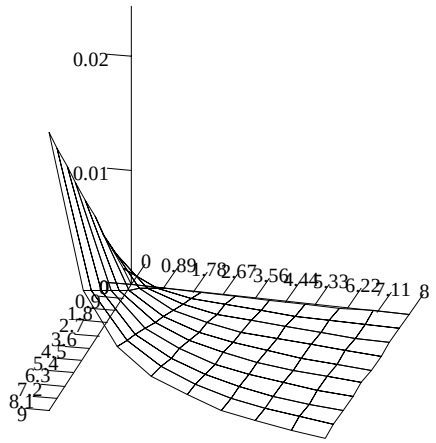
DEFINITION 4. The level surface Σ

$$\Sigma = \{ (x, y, z) \in \mathbf{R}^3 : \frac{xy}{z(1-y)} = \frac{\Delta}{M - \Delta} \}$$

is called the *effectiveness surface* of *IR* (corresponding to *q*).

The figure below shows a plot of the effectiveness surface.

The vertical axis corresponds to fallout, the axis to its right to precision, and the third axis corresponds to recall. It can nicely be seen that fallout increases when precision is low and recall is high. The lower on the surface, the higher precision and/or recall. The shape of the effectiveness surface is the same for every IR system hence it is a typical surface. Its specific position in space varies with the total number of documents in the IR system and the total number of relevant documents given a query. The surface remains the same for the same query, and repeated relevance feedbacks for the same query means a walk on this surface (which plays the role of a constraint in the optimisation of effectiveness).



Effectiveness surface E.

4. Relevance feedback as a Diophantine structure

Let *q* denote a query. In Probabilistic *IR* (*PIR*), the set $\Re(q)$ of retrieved documents $d \in D$ in response to query *q* contains those documents whose conditional probability of relevance $P(R|(q, d))$ exceeds that of nonrelevance $P(I|(q, d))$ – Bayes' Decision Rule – and a real valued threshold (or cut off) $\tau \geq 0$, i.e. $\Re(q) = \{ d \in D: P(R|(q, d)) \geq P(I|(q, d)), P(R|(q, d)) \geq \tau \}$.

A basis to estimate the probabilities $P(.|(q, d))$ is offered by Bayes' Formula:

$$P(\bullet|(q, d)) = \frac{P((q, d)|\bullet)\, P(\bullet)}{P(q, d)}$$

where the symbol $\bullet$ stands for relevance *R* or irrelevance *I*.

Bayes' Formula requires that for the estimation of $P(\bullet|(q, d))$ an initial set $\Re_0(q)$ be known first, based on which $P((q, d)|\bullet)$ can be estimated. The denominator $P(q, d)$ simplifies, $P(\bullet)$ is

the a priori probability of criterion . in general and is constant. The estimation of $P(\bullet|(q, d))$ can be iterated using each time the previous $\Re(q)$ to re-estimate (relevance feedback) the probabilities $P((q, d)|\bullet)$.

We show now the following connection between *PIR* and a Diophantine structure.

THEOREM 1. *Repeatedly applying PIR yields a Diophantine structure.*

Proof. Given a query q. An initial set $\Re_0(q)$ of retrieved documents is obtained first. *PIR* is repeatedly applied in consecutive steps s = 1, 2, ... At any step s, the set $\Re_{s-1}(q)$ of the previous step is used to estimate the probabilities $P((q, d)|\bullet)$ based on which the probabilities $P(\bullet|(q, d))$ can be calculated – using Bayes' Formula – and a new set $\Re_s(q)$ of retrieved documents is obtained. Let $f(x, y)$ mean the newely retrieved set of documents $\Re_s(q)$ at step s, where x is an integer variable corresponding to query q and y is an integer variable symbolising step s when probabilities $P(\bullet|(q, d))$ are computed. Let the process of calculating, based on relevance feedback, the new probabilities $P(\bullet|(q, d))$ and of retrieving a new set $\Re_{s+1}(q)$ of documents, at step s+1, be represented by a function $\beta(x, y, f(x, y))$. One can consider a series $\Re_0(q)$, $\Re_1(q)$, $\Re_2(q)$, ..., $\Re_s(q)$, ... of retrieved documents. Thus, using the above introduced functions, one can define function f recursively as follows: $f(x, 0) = \alpha(x)$ and $f(x, y + 1) = \beta(x, y, f(x, y))$ with the following meaning: at the initial step s = 0, i.e. $f(x, 0)$, an initial set $\Re_0(q)$ is retrieved (using e.g. a vector or interaction or another method), represented by $\alpha(x)$; then, at every next step s + 1, a new set $\Re_{s+1}(q)$ is obtained, i.e. $f(x, y + 1)$, after repeatedly computing, based on relevance feedback using the previous $\Re_s(q)$, the probabilities $P(\bullet|(q, d))$ and perfoming a retrieval opeartion again, i.e. $\beta(x, y, f(x, y))$. Because, formally, function f is recursively defined (primitive recursive function), the series $\Re_0(q)$, $\Re_1(q)$, $\Re_2(q)$, ..., $\Re_s(q)$, ...

forms a recursively enumerable (r.e.) set (relative to the power set $\wp(D)$ where D denotes the set of dcouments to be searched), and, as such, it is a Diophantine set.

The function f, being recursive, is computable, hence it has a fixed point (Rogers Fixed Point Theorem, Philips, 1992). This means that there exists an index s such that the same $\Re_s$ is obtained in a next step as that of the previous one, i.e. there is not any improvement and thus this can be inerpreted as an optimal situation. If one assigns now to the sets $\Re_s$ points on a the effectiveness surface, a fixed point corresponds to an (local/global) optimum (minimum, maximum) on this surface.

5. Optimal Information Retrieval

Ideally, an *IR* should be such that $\varphi = 0$ and $\rho = 1$. Because $\varphi = 0$ implies $\pi = 1$, the following two optimalities can be defined.

DEFINITION 6. An *IR* is called ρ-*optimal* if $\rho = 1$, and φ-*optimal* if $\varphi = 0$.

It is easy to see that in a ρ-optimal *IR* we have $\pi < 1$, whereas in a φ-optimal *IR* $\rho < 1$. Effectiveness (or performance) is now defined as a relative position, e.g. Euclidean distance or a cosine, to the ideal $I^* = (\varphi^*, \rho^*, \pi^*) = (0, 1, 1)$ as follows.

DEFINITION 7. *Effectiveness* (*performance*) ε of an *IR* is defined as the Euclidean distance between a current point (φ, ρ, π) and the ideal $I^* = (0, 1, 1)$:

$$\varepsilon = \left(\varphi^2 + (1 - \pi)^2 + (1 - \rho)^2\right)^{1/2}$$

The smaller ε the more effective *IR*. It is easy to see that $\varepsilon = 1 - \rho$ for φ-optimality, and $\varepsilon = (\varphi^2 + (1 - \pi)^2)^{1/2}$ for ρ-optimality. The concept of a global optimality is now defined as follows (its existence follows from Theorem 1):

DEFINITION 8. An *optimal IR* is one with φ, π and ρ such that $\min_{\varphi,,\pi,\rho\in\Sigma}\varepsilon$.

In other words, an *IR* is optimal if its recall, precision and fallout are a solution of the following nonlinear minimization problem with constraints:

$$\min\ (\varphi^2 + (1 - \pi)^2 + (1 - \rho)^2)^{1/2}$$

subject to the following constraints:

$$\frac{\varphi\pi}{\rho(1-\pi)} = \frac{\Delta}{M - \Delta} \qquad 0 < \rho \le 1 \qquad 0 \le \pi \le 1$$

Alternatively, instead of the Euclidean distance, other measure for ε may also be used, e.g. the cosine of the angle between the ideal $\boldsymbol{I}^* = (0, 1, 1)$ and the current vector $\mathbf{o} = (\varphi, \pi, \rho)$:

$$\varepsilon = \frac{(\boldsymbol{I}^*, \mathbf{o})}{\|\boldsymbol{I}^*\|\,\|\mathbf{o}\|} = \frac{\pi + \rho}{2\cdot(\varphi^2 + \pi^2 + \rho^2)^{1/2}}$$

In this case, the higher ε, the more effective *IR* (the smaller the angle). The concept of a global optimality in this case is defined as follows. An *optimal IR* is one with φ, π and ρ such that $\max_{\varphi,\pi,\rho\in\Sigma}\varepsilon$.

In other words, an *IR* is optimal if its recall, precision and fallout are a solution of the following nonlinear maximization problem with constraints:

$$\max \frac{\pi + \rho}{2\cdot(\varphi^2 + \pi^2 + \rho^2)^{1/2}}$$

subject to the following constraints:

$$\frac{\varphi\pi}{\rho(1-\pi)} = \frac{\Delta}{M-\Delta} \qquad 0 < \rho \le 1 \qquad 0 \le \pi \le 1$$

Solutions for both problems are as follows (using MathCAD 8.01 Plus Professional). Both methods give, practically, the same global optimum.

a) **Euclidean effectiveness** :

$\varepsilon(\phi,\rho,\Pi) := \sqrt{\phi^2 + (1-\Pi^2) + (1-\rho^2)}$ Let $\Delta := 500$ $M := 500000$

Guess values (near to global optimum): $\phi := 0.1$ $\Pi := 0.9$ $\rho := 0.9$

Minimization with constraints:

Given

$\frac{\phi\cdot\Pi}{\rho\cdot(1-\Pi)} = \frac{\Delta}{M-\Delta}$ $\rho > 0$ $\rho \le 1$ $\Pi \ge 0$ $\Pi \le 1$

$$\text{Minimize}(\varepsilon,\phi,\rho,\Pi) = \begin{bmatrix} 1.275\cdot10^{-7} \\ 1 \\ 1 \end{bmatrix}$$

b) **Cosine angle effectiveness:**

$$\varepsilon(\phi,\rho,\Pi) := \frac{\Pi+\rho}{2\cdot\sqrt{\phi^2+\Pi^2+\rho^2}}$$

Maximization with constraints:

Given

$\frac{\phi\cdot\Pi}{\rho\cdot(1-\Pi)} = \frac{\Delta}{M-\Delta}$ $\rho > 0$ $\rho \le 1$ $\Pi \ge 0$ $\Pi \le 1$

$$\text{Maximize}(\varepsilon,\phi,\rho,\Pi) = \begin{bmatrix} 7.625\cdot10^{-7} \\ 1 \\ 0.999 \end{bmatrix}$$

6. Conclusion

It was shown that a typical surface can be constructed on which each point corresponds to a three-tuple (recall, precision, fallout). The name effectiveness surface was suggested for it. A series of repeatedly applied relevance feedbacks corresponds to a sequence of points on this surface. It was shown that the corresponding series of sets of retrieved documents forms a Diophantine set which, as a recursive structure, has a fixed point. This point may be interpreted as being an optimal point for optimal effectiveness values. These values can be calculated, they being the solutions of a constrained nonlinear optimisation mathematical problem. In other words, relevance effectiveness enhancement is formulated as a constrained nonlinear optimisation problem controlled by an effectiveness surface. Thus, relevance feedback, as a means to enhance effectiveness, is equivalent to a mathematical process of minimizing (maximizing) a nonlinear function subject to nonlinear constraints.

The paper also shows a connection between IR and Recursion Theory as an abstract mathematical structure.

References

Allan, J. (1996) Incremental Relevance Feedback. In SIGIR '96 *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*. Zurich, Switzerland, 270-278.

Belkin, N.J. et al. (1996) Using relevance feedback and ranking in interactive searching. In Harman, D. (ed.) *TREC-4 Proceedings of Fourth Text Retrieval Conference*. Washington, D.C., 181-209.

Belkin, N.J. & Koenemann, J. (1996) A case for interaction: A study of interac-tive information retrieval behavior and effectiveness. In *Proceddings of the ACM SIG CHI Conference on Human Factors in Computing Systems*, New York, 205-212.

Buckland, M. & Gey, F. (1994) The Relationship Between Recall and Precision. *Journal of the American Society for Information Science*, **45(1)**: 12-19.

JASIS (1994) Special topic issue: Relevance research. *Journal of the American Society for Information Science*, 45.

Mizzaro, S. (1997) Relevance: The Whole History. *Journal of the American Society of Information Science*. 48, 810-832.

Philips, I.C.C. (1992) Recursion Theory. In: Abramsky, S. & Gabbay, D.M. & Maibaum, T.S.E. (eds.) *Handbook of Logic in Computer Science.* Vol. 1, Oxford Science Publications, Clarenden Press.