Mathematica Pannonica
11/1 (2000), 137 – 153

FOUNDATION OF INFORMATION RETRIEVAL

Sándor **Dominich**

Department of Computer Science, University of Veszprém, 8200 Veszprém, Egyetem u. 10, Hungary

Dedicated to Professor Gino Tironi on his 60th birthday

Received: January 1999

MSC 1991: 68 P 20

Keywords: Information retrieval, mathematical foundation.

Abstract: A mathematical foundation and theory for Information Retrieval (*IR*) is created (for the first time). A unified definition for the classical (vector, probabilistic) models is given. Abstract mathematical structures, characterising retrieval, are revealed based on which mathematical theories are created. Mathematical definition for relevance effectiveness is given. All particular mathematical results achieved in *IR* so far can be integrated into this new theory. Moreover, it is shown that the classical models of *IR* are a special case of Interaction *IR* (I^2R) — beside that of Associative *IR* (AI^2R). Thus, *IR* becomes a mathematical discipline (too), formally comparable to other scientific disciplines, researchable by rigorous mathematical means, methodically accessible by and teachable to a more diverse scientific audience. Also, links between *IR* and modern mathematical disciplines are formally created.

1. Introduction

E-mail address: dominich@dcs.vein.hu

This research was partly supported by research grant Pro Renovanda Tud. 98/31, and the writing of this paper has been begun while the author was with the Buckinghamshire Chilterns University College, United Kingdom (1996–1997).

Generalizations ([16], [33], [38]) of the probabilistic *IR* yielded the idea of creating a unified mathematical foundation and theory of *IR*, the importance of which consists in that making full use of rapidly developing technological advances (Internet, CD-ROM industry) can only be expected on the basis of a unified and consistent mathematical theory. Because these generalizations shifted the interpretation of conditional probabilities (which play a central role in probabilistic *IR*) towards logic, *logic IR* models were elaborated. They constitute a new model type of *IR*, and, on the other hand, aimed at containing the classical models of *IR*, too, as special cases ([5], [4], [34], [18]) but without success.

The present paper realizes this idea (for the first time), on a mathematical basis. All parts, apart from Part 4, are new and original.

Firstly, the concepts of classical *IR* models (vector, probabilistic) are recalled and mathematically defined to fix the ideas (Part 2). Then, a unified definition is formulated, and it is shown that the two classical models are two special cases of it (Part 3). Part 4 recalls the concepts of relevance effectiveness measures which are used later on.

The next part, Part 5, creates the basis for a mathematical theory of the classical vector model. A naturally arising link to metric spaces and metric induced topological spaces is revealed, and it is shown that retrieval means inducing a topology. This is the case in typical vector models (and thus usual database searching as well). Also, the link to nonmetric topologies is justified, a point to which recent research links. In this case, too, retrieval means defining a topology (using open sets). A concept of optimal relevance effectiveness is defined as finding all sequences of documents that converge (either in a metric or vicinities-based sense) to a query (Cauchy-sequences).

Part 6 creates the basis for a mathematical theory of the probabilistic model. It is shown that the main technique used for repeatedly enhanced retrieval, namely relevance feedback, yields a Diophantine set (of retrieved documents). Thus, a fixed point exists which yields a mathematical formulation of optimal relevance effectiveness as a contrained nonlinear optimisation problem controlled by a very specific surface.

Part 7 shows that the unified definition of the classical models is a special case of a more general (and nonclassical) model of *IR*: Interaction *IR* (I^2R). It is also shown that there exists another special case of *IR* as well called Associative *IR* (AI^2R) in which retrieval means

defining a matroid.

Because *IR*, due to this new, unified and consistent mathematical theory, now has its own mathematical individuality, it can be formally subjected to various mathematical considerations. Examples are shown in Part 8: comparisons with decision theory and situation theory, and computational complexity aspects.

Proofs are only given for theorems, lemmas, etc. that are particularly relevant to this new theory of *IR*.

2. The classical models of *IR*

The basic models can be defined as follows (based on [32], [27], [28], [39], [35]):

2.1. *Definition of Vector IR (DVIR).* Given a nonempty finite set D, a threshold value $\tau \in \mathbb{R}$ and two mappings $\sigma : D \times D \to [0;1]$ and $\Re : D \to \wp(D)$ such that: i) $0 \leq \sigma(a,b) \leq 1$, $\forall a, b \in D$ (normalization); ii) $\sigma(a,b) = \sigma(b,a)$, $\forall a, b \in D$ (commutativity); iii) $a = b \Rightarrow \sigma(a,b) = 1$, $\forall a, b \in D$ (reflexivity); iv) $\Re(q) = \{d \in D : \sigma(d,q) > \tau\}$ (retrieval). Then the 3-tuple $\langle D, \sigma, \Re \rangle$ is called a *Vector IR*.

σ is called a *similarity*. See [19], [13] fore more on this.

2.2. *Definition of Probabilistic IR (DPIR).* Given a nonempty finite set D, a cut-off value $\alpha \in \mathbb{R}$, a probability measure P over $\wp(D \times \times D)$. Let $R_q \subseteq D \times D$ and $I_q \subseteq D \times D$ be two binary relations, $q \in D$, and $\Re : D \to \wp(D)$, $\Re(q) = \{d \in D : P(R_q \mid \{(q,d)\}) \geq \geq P(I_q \mid \{(q,d)\}) \wedge P(R_q \mid \{(q,d)\}) > \alpha\}$ a mapping (retrieval). Then the 3-tuple $\langle D, P, \Re \rangle$ is called a *Probabilistic IR*.

$P(R_q \mid \{(q,d)\}) \geq P(I_q \mid \{(q,d)\})$ is called *Bayes' Decision Rule.*

2.3. Further models are defined using the basic models.

3. Unified definition of the basic models

Let

$T = \{t_1, t_2, \ldots, t_k \ldots t_N\}$ be a set of *identifiers*, $N \geq 1$;

$O = \{o_1, o_2 \ldots o_u, \ldots, o_U\}$ a set of *objects*, $U \geq 1$;

$(D_j \in \wp(O))_{j \in J = \{1,2,\ldots,M\}}$ a family of *object clusters*, $M \geq 2$;

$D = \{\tilde{o}_j : \forall j \in J\}$ a set of *documents*, $\tilde{o}_j = \{(t_k, \mu_{\tilde{o}_j}(t_k)) : t_k \in \in T,\ k = 1, 2, \ldots, N\}$ is a fuzzy set and *cluster representative* of D_j, $\mu_{\tilde{o}_j} : T \to S \subseteq [0;1]$, $\forall j \in J$;

$A = \{\tilde{a}_1, \tilde{a}_2 \ldots \tilde{a}_i \ldots \tilde{a}_C\}$ be a set of *criteria*, $C \geq 1$ where $\tilde{a}_i = \{((q, \tilde{o}_j), \mu_{\tilde{a}_i}(q, \tilde{o}_j)) : \tilde{o}_j \in D,\ j = 1, 2, \ldots, M\}$, is a fuzzy relation, $i = 1, 2, \ldots, C$, $\mu_{\tilde{a}_i} : D \times D \to [0; 1]$, $q \in D$ arbitrary fixed;

$a_{\alpha_i} = \{\tilde{o} \in D : \mu_{\tilde{a}_i}(q, \tilde{o}) > \alpha_i\}$, $i = 1, 2, \ldots, C$, an α_i-cut of $\tilde{a}_i$, $0 \leq \alpha_i < +\infty$, $q \in D$ arbitrary fixed (a_{α_i} is used rather than $a_{q\alpha_i}$ to avoid using too many indices);

$\Re : D \to \wp(D)$ a mapping called *retrieval*.

All sets are finite. See also [20] for more on fuzzy techniques used in *IR*. All fuzzy sets are normalised.

3.1. A unified definition of the classical models is as follows.

Definition 1. A 2-tuple $\langle D, \Re \rangle$ is called a *Classical Information Retrieval* (*CIR*) if the following properties **P1** and **P2** hold: **P1.** $q = \tilde{o} \Rightarrow \mu_{\tilde{a}_i}(q, \tilde{o}) = 1$, $\forall q, \tilde{o} \in D$, $i = 1, 2, \ldots, C$; **P2.** $\Re(q) = \{\tilde{o} : \mu_{\tilde{a}_i}(q, \tilde{o}) = \max_{k=1,\ldots,C} \mu_{\tilde{a}_k}(q, \tilde{o})\} \cap a_{\alpha_i}$, i arbitrary fixed, $k = 1, \ldots, C$.

3.2. A special case of *CIR* is as follows.

Definition 2. *Similarity Information Retrieval* (*SIR*) is a *CIR* $\langle D, \Re \rangle$ if properties **S1** and **S2** hold: **S1.** $C = 1$; **S2.** $\mu_{\tilde{a}_i}(q, \tilde{o}) = \mu_{\tilde{a}_i}(\tilde{o}, q)$, $\forall \tilde{o}$, $q \in D$ (commutativity).

It is now shown that the classical vector model is a special case of *CIR*.

Theorem 1. *DVIR and SIR are equivalent.*

Proof. "$\Leftarrow$". $C = 1$ by **S1**. Let $\tilde{a}_1$ be this criterion and called 'similar'. By **S2**: $\mu(q, \tilde{o}) = \mu(\tilde{o}, q)$, $\forall \tilde{o}$, $q \in D$. By **P1**: $q = \tilde{o} \Rightarrow \mu(q, \tilde{o}) = 1$, $\forall \tilde{o}$, $q \in D$. μ is normalized: $0 \leq \mu(q, \tilde{o}) \leq 1$, $\forall \tilde{o}$, $q \in D$. Hence μ satisfies i)–iii) of *DVIR*. **P2** becomes: $\Re(q) = \{\tilde{o} : \mu(q, \tilde{o}) > \alpha\}$. "$\Rightarrow$" σ is a degree to which b is similar to a, thus $\sigma(a, b)$ is conceived as $\mu_{\tilde{a}_1}(a, b)$ where $\tilde{a}_1$ denotes 'similar'. This is equivalent to **S1**. Condition i) is equivalent to normalization, ii) to **S2**, whereas iii) to **P1**. The retrieval condition becomes: $\Re(q) = \{d \in D : \sigma(d, q) > \tau\} = \{d \in D : \mu_{\tilde{a}_1}(d, q) = \max \mu_{\tilde{a}_1}(d, q)\} \cap a_\tau$. $\Diamond$

Two special cases of *SIR* are as follows:

Definition 3. *Binary SIR* (*BSIR*) is a *SIR* with $S = \{0, 1\}$.

Definition 4. *Non-Binary SIR* (*NBSIR*) is a *SIR* where S = [0; 1].

3.3. Another special case of *CIR* is as follows:

Definition 5. *Probabilistic IR* (*PIR*) is a *CIR* $\langle D, \Re \rangle$ in which the following property **P** holds: $C = 2$.

It is now shown that the classical probabilistic model is a special case of *CIR* by showing that it is equivalent to *PIR*. But the following lemma is needed first.

Lemma 1. *Given* T, D *and* $\tilde{a}_i$, $i = 1, 2$. *Let* P *be a probability measure and* $p_{kj}^{(i)} = P(X_k = \mu_{\tilde{o}_j}(t_k))$, $i = 1, 2$. *If i)* $\mu_{\tilde{a}_i}(q, \tilde{o}_j) = \sum_{k=1}^{N} \log(p_{kj}^{(1)} / p_{kj}^{(2)})$, $i = 1, 2$; ii) *identifier occurrences are independent;* iii) *the two criteria are disjoint; then* 1) *The degree of compatibility of an object with a criterion is directly proportional to the conditional probability of that criterion given the object, i.e.* $\mu_{\tilde{a}_1}(q, \tilde{o}_j) \geq \mu_{\tilde{a}_1}(q, \tilde{o}_s) \Leftrightarrow P(\tilde{a}_1 \mid \tilde{o}_j) \geq \geq P(\tilde{a}_1 \mid \tilde{o}_s)$, *and* 2) $\mu_{\tilde{a}_1}(q, \tilde{o}_j) \geq \mu_{\tilde{a}_2}(q, \tilde{o}_j) \Leftrightarrow P(\tilde{a}_1 \mid \tilde{o}_j) \geq P(\tilde{a}_2 \mid \tilde{o}_j)$.

Proof. 1) On can write $\mu_{\tilde{a}_i}(q, \tilde{o}_j) = \sum_{k=1}^{N} \log(p_{kj}^{(1)} / p_{kj}^{(2)}) = \log \prod_{k=1}^{N} p_{kj}^{(1)} / p_{kj}^{(2)}$. Any object $\tilde{o}_j$ is a compound event. Thus a probability $P_i((q, \tilde{o}_j)) = = P_i(\tilde{o}_j) = \prod_{k=1}^{N} p_{kj}^{(i)}$, $i = 1, 2$, can be constructed: $P_i(\tilde{o}_j)$ is a conditional probability of $\tilde{o}_j$ given relevance $\tilde{a}_1$ or non-relevance $\tilde{a}_2$, and can thus be denoted by $P_1(\tilde{o}_j) = P(\tilde{o}_j \mid \tilde{a}_1)$ and $P_2(\tilde{o}_j) = P(\tilde{o}_j \mid \tilde{a}_2)$, respectively. $P_i(\tilde{o}_j)$ is unique for $\tilde{o}_j$. Thus:

$$\frac{P_1(\tilde{o}_j)}{P_2(\tilde{o}_j)} = \frac{P(\tilde{o}_j \mid \tilde{a}_1)}{P(\tilde{o}_j \mid \tilde{a}_2)} = \frac{\prod_{k=1}^{N} p_{kj}^{(1)}}{\prod_{k=1}^{N} p_{kj}^{(2)}} = \prod_{k=1}^{N} \frac{p_{kj}^{(1)}}{p_{kj}^{(2)}}.$$

Hence

$$\mu_{\tilde{a}_1}(q, \tilde{o}_j) \geq \mu_{\tilde{a}_1}(q, \tilde{o}_s) \Leftrightarrow \frac{P(\tilde{o}_j \mid \tilde{a}_1)}{P(\tilde{o}_j \mid \tilde{a}_2)} \geq \frac{P(\tilde{o}_s \mid \tilde{a}_1)}{P(\tilde{o}_s \mid \tilde{a}_2)}.$$

It follows that

$$\frac{P(\tilde{o}_j \mid \tilde{a}_1)}{P(\tilde{o}_j \mid \tilde{a}_2)} - \frac{P(\tilde{o}_s \mid \tilde{a}_1)}{P(\tilde{o}_s \mid \tilde{a}_2)} \geq 0 \Leftrightarrow$$

$$\Leftrightarrow P(\tilde{o}_j \mid \tilde{a}_1) P(\tilde{o}_s \mid \tilde{a}_2) - P(\tilde{o}_s \mid \tilde{a}_1) P(\tilde{o}_j \mid \tilde{a}_2) \geq 0.$$

By Bayes' Formula:

$$P(\tilde{a}_1 \mid \tilde{o}_j) = \frac{P(\tilde{o}_j \mid \tilde{a}_1) P(\tilde{a}_1)}{P(\tilde{o}_j \mid \tilde{a}_1) P(\tilde{a}_1) + P(\tilde{o}_j \mid \tilde{a}_2) P(\tilde{a}_2)}$$

$$P(\tilde{a}_1 \mid \tilde{o}_s) = \frac{P(\tilde{o}_s \mid \tilde{a}_1) P(\tilde{a}_1)}{P(\tilde{o}_s \mid \tilde{a}_1) P(\tilde{a}_1) + P(\tilde{o}_s \mid \tilde{a}_2) P(\tilde{a}_2)}.$$

Hence $P(\tilde{a}_1 \mid \tilde{o}_j) - P(\tilde{a}_1 \mid \tilde{o}_s) \geq 0$, and thus $\mu_{\tilde{a}_1}(q, \tilde{o}_j) \geq \mu_{\tilde{a}_1}(q, \tilde{o}_s) \Leftrightarrow \Leftrightarrow P(\tilde{a}_1 \mid \tilde{o}_j) \geq P(\tilde{a}_1 \mid \tilde{o}_s)$.

2) The condition $\mu_{\tilde{a}_1}(q, \tilde{o}_j) \geq \mu_{\tilde{a}_2}(q, \tilde{o}_j)$ rewrites $\mu_{\tilde{a}_1}(q, \tilde{o}_j) = = \mu_{\tilde{a}_2}(q, \tilde{o}_j) \geq c$ (some threshold c). That is

$$\log \prod_{k=1}^{N} \frac{p_{kj}^{(1)}}{p_{kj}^{(2)}} \geq c \Leftrightarrow \frac{P(\tilde{o}_j \mid \tilde{a}_1)}{P(\tilde{o}_j \mid \tilde{a}_2)} \geq C \geq 1.$$

Because b can be chosen such that $b^c \geq P(\tilde{a}_2)/P(\tilde{a}_1)$ it follows that

$$\frac{P(\tilde{o}_j \mid \tilde{a}_1)}{P(\tilde{o}_j \mid \tilde{a}_2)} \geq \frac{P(\tilde{a}_2)}{P(\tilde{a}_1)}.$$

Using the Bayes' Formula one gets:

$$\frac{P(\tilde{o}_j \mid \tilde{a}_1)P(\tilde{a}_1)}{P(\tilde{o}_j)} \geq \frac{P(\tilde{o}_j \mid \tilde{a}_2)P(\tilde{a}_2)}{P(\tilde{o}_j)}$$

which is equivalent to $P(\tilde{a}_1 \mid \tilde{o}_j) \geq P(\tilde{a}_2 \mid \tilde{o}_j)$ $\Diamond$

The assumptions in Lemma 1 are typical in the probabilistic model. A special case of this lemma can be found in [39] where it is shown that this is an optimal retrieval. Based on Lemma 1, it is now shown that the classical probabilistic model is a special case of *CIR* by showing that it is equivalent to *PIR*.

Theorem 2. *PIR and DPIR are equivalent.*

Proof. **P2** becomes: $\Re(q) = \{\tilde{o} : \mu_{\tilde{a}_i}(q, \tilde{o}) = \max_{k=1,\ldots,C} \mu_{\tilde{a}_k}(q, \tilde{o})\} \cap \cap\, a_{\alpha_i} = \{\tilde{o} : \mu_{\tilde{a}_i}(q, \tilde{o}) \geq \mu_{\tilde{a}_j}(q, \tilde{o}), j = i + (-1)^{i+1}, \mu_{\tilde{a}_i}(q, \tilde{o}) > \alpha_i\}$. This rewrites as (Lemma 1): $\Re(q) = \{\tilde{o} : P(\tilde{a}_i \mid \tilde{o}) \geq P(\tilde{a}_j \mid \tilde{o}), j = i + + (-1)^{i+1}, P(\tilde{a}_i \mid \tilde{o}) > \alpha_i\}$ which is exactly the retrieval condition in *DPIR*. **P1** ensures that reflexivity holds too. $\Diamond$

Two special cases of *PIR* are as follows.

Definition 6. *Binary Probabilistic IR* (*BPIR*) is a *PIR* with $S = \{0, 1\}$.

Definition 7. *Non-Binary Probabilistic IR* (*NBPIR*) is a *PIR* with $S = [0; 1]$.

3.4. As a noteworthy property, it can be shown that *SIR* may be conceived as a special case of *PIR*.

Theorem 3. *Let* $\langle D, \Re \rangle$ *be a PIR. If* $\tilde{a}_1 = \tilde{a}_2$ *and* $\mu_{\tilde{a}_1}(q, d) = \mu_{\tilde{a}_1}(d, q)$ $\forall d, q$ *then PIR is a SIR.* $\Diamond$

4. Relevance effectiveness measures

Given a *CIR* $\langle D, T\Re \rangle$. Let $|D| = M$, $|T\Re(q)| = \Delta \neq 0$ (the elements of $T\Re(q)$ are said to be relevant to q), $M > \Delta$. Let $\langle D, \Re \rangle$ be another *CIR*, and let $|\Re(q)| = \kappa \neq 0$, $|\{d \in \Re(q)\colon d \in T\Re(q)\}| = \alpha$. The usual effectiveness measures are:

$$\text{Recall: } \rho = \frac{\alpha}{\Delta}, \quad \text{Precision: } \pi = \frac{\alpha}{\kappa}, \quad \text{Fallout: } \varphi = \frac{\kappa - \alpha}{M - \Delta}.$$

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

Proof. $\alpha = \rho\Delta = \pi\kappa \Rightarrow \rho/\pi = \kappa/\Delta$, $(1/\Delta = \text{constant})$. $\Diamond$

A generalization — using polynomials — of Prop. 1 is found in [2] where the same result is obtained based on empirical axioms.

5. Basis for a mathematical theory of *SIR*

It is convenient to reformulate the concept of *SIR* as follows.

Definition 8. Let D be a set of objects and σ a similarity on D. A *similarity space* (or *σ-space*) on D is a 2-tuple $\langle D, \sigma\rangle$.

Definition 9. Let $\langle D, \sigma\rangle$ be a σ-space. *Similarity Information Retrieval* (*SIR*) on D is a 3-tuple $\langle D, \Re, \sigma\rangle$ where $\Re(d) = \{x \in D : \sigma(d, x) > \tau\}$.

It is relatively easy to see that the following theorems hold.

Theorem 4. *Let $\langle E, \mu\rangle$ be a pseudometric space. Then $\langle E, 1 - \mu/ / \max_E \mu\rangle$ is a σ-space.* $\Diamond$

Definition 10. Let $\langle E, \mu\rangle$ be a pseudometric space. Then $\langle E, 1 - \mu/ / \max_E \mu\rangle$ is the *σ-space induced* on E *by pseudometric* μ.

Theorem 5. *Let $\langle E, \mu\rangle$ be a pseudometric space. Then the induced topological space is a SIR on E.* $\Diamond$

Theorem 6. *Let $\langle E, \mu\rangle$ be a pseudometric space. Then the relation $\tilde{}$ defined as $x \tilde{} y \Leftrightarrow \mu(x, y) = 0$, $\forall x, y \in E$, is an equivalence relation on E.* $\Diamond$

Let E^* denote the set of equivalence classes.

Theorem 7. *Let $\langle E, \mu\rangle$ be a pseudometric space. The space $\langle E^*, \mu^* = \mu\rangle$ is a metric space with $\mu^*(A, B) = \mu(x, y)$, $A, B \in E^*$, $x \in A, y \in B$.* $\Diamond$

Corollary 1. *The Hausdorff space induced by metric μ^* is a SIR.* $\Diamond$

Corollary 2. *Let $\langle E, \sigma\rangle$ be a σ-space. If $\delta = 1 - \sigma$ is a pseudometric/metric on E, then the induced topological space/Hausdorff space on E and SIR are equivalent.* $\Diamond$

In [12] it is shown that $BSIR = \langle D, \sigma = \text{Dice's Coefficient}\rangle$, which is a typical *SIR*, satisfies the theorems above. Another typical *SIR*, namely $\langle D, \sigma = \text{binary/nonbinary Cosine Measure}\rangle$, also satisfies them, but just in general. Therefore non-metrical topologies on σ-spaces are of interest, too; they are investigated in [15], [14].

A definition of an optimality for relevance effectiveness is as follows. $\pi = 1$ is equivalent to $\Re(d)$ being a (correspondingly defined) Cauchy-sequence. Hence the ideal case $\pi = 1$, $\rho = 1$ yields:

Definition 11. An $SIR = \langle D, \Re, \sigma\rangle$ is *optimal* if $\Re(d)$ is the maximal subspace of D consisting of Cauchy-sequences.

6. Basis for a mathematical theory of *PIR*

Let us denote Bayes' Decision Rule by **B**: $P(\tilde{a}_1 \mid \tilde{o}_s) \geq P(\tilde{a}_2 \mid \tilde{o}_s)$. Or simply $j = 1$ (relevant), 2 (non-relevant) for criterion $\tilde{a}_j$, (document) d for object $\tilde{o}_s$, query q for object q and $P_q(j \mid d)$ denote $P(\tilde{a}_j \mid (q, \tilde{o}_j))$. The probabilities $P_q(j \mid d)$ are typically estimated using Bayes' Formula (see [21] for another method):

$$P_q(j \mid d) = \frac{P_q(d \mid j) P_q(j)}{P_q(d)}.$$

The following equivalence holds (the cut-off value may be neglected here): $\Re(q) = \{d : \mu_{\tilde{a}_1}(q,d) \geq \mu_{\tilde{a}_2}(q,d)\} \Leftrightarrow \mathbf{B}(\Re(q)) : P_q(1 \mid d) \geq \geq P_q(2 \mid d)$ where the dot in $\mathbf{B}(.)$ denotes the set whose elements are decided on. Bayes' Formula requires that for the estimation of $P_q(j \mid d)$ an initial set $\Re_0(q)$ be known first, based on which $P_q(d \mid j)$ can be estimated. Then, after an initial step, *PIR* can be iterated using each time the previous $\Re(q)$ to re-estimate (relevance feedback) the probabilities $P_q(d \mid j)$.

Theorem 8. *Let $\langle D, \Re \rangle$ be a PIR, $\Re_0(q)$ an initial set of retrieved objects. Then repeatedly applying PIR yields a Diophantine set.*

Proof. Given a query q. An initial set $\Re_0(q)$ of retrieved documents is obtained first. *PIR* is repeatedly applied in consecutive steps $s = = 1, 2, \ldots$. At any step s, the set $\Re_{s-1}(q)$ of the previous step is used to estimate the probabilities $P_q(d \mid j)$ based on which the probabilities $P_q(j \mid d)$ 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 $P_q(j \mid d)$ are computed. Let the process of calculating, based on relevance feedback, the new probabilities $P_q(j \mid d)$ 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 $\Re_0(q), \Re_1(q), \Re_2(q), \ldots, \Re_s(q), \ldots$ of retrieved documents. Thus, using the above introduced functions, one can define a 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 $P_q(j \mid d)$ and

performing a retrieval operation again, i.e. $\beta(x, y, f(x, y))$. Because, formally, function f is recursively defined (primitive recursive function), the series $\Re_0(q), \Re_1(q), \Re_2(q), \ldots, \Re_s(q), \ldots$ forms a recursively enumerable (r.e.) set (relative to the power set $\wp(D)$ where D denotes the set of documents to be searched), and, as such, it is a Diophantine set. $\Diamond$

The function f, being recursive, is computable, hence — based on Rogers Fixed Point Theorem ([26]) — it has a fixed point. This means that there exists an index s such that the same $\Re_s$ is obtained (in a next step) when rule **B** is applied upon $\Re_s$, i.e. **B** $(\Re_s) = \Re_s$. If one assignes points to the sets $\Re_s$ on a surface, a fixed point corresponds to an optimum. An important property of recall, precision and fallout is as follows:

Theorem 9.

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

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

This makes it possible to define the following specific surface.

Definition 12. The level surface

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

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

Ideally, an *IR* should be such that $\varphi = 0$, $\pi = 1$ and $\rho = 1$. Thus:

Definition 13. $\epsilon = \sqrt{\varphi^2 + (1-\pi)^2 + (1-\rho)^2}$ is called the *effectiveness* of *PIR*.

Definition 14. A *PIR* is *optimal* if φ, π and ρ minimise ϵ.

In other words (constrained nonlinear optimisation):

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

$$\text{subject to } \frac{\varphi\pi}{\rho(1-\pi)} = \frac{\Delta}{M-\Delta},\ 0 \le \rho \le 1,\ 0 \le \pi \le 1.$$

Alternatively, the cosine of the angle between the ideal vector $\mathbf{v} = (0, 1, 1)$ and the actual vector $\mathbf{o} = (\varphi, \pi, \rho)$ can also be used:

$$\epsilon = \frac{(\mathbf{v}, \mathbf{o})}{\|\mathbf{v}\| \, \|\mathbf{o}\|} = \frac{\pi + \rho}{\sqrt{2} \cdot \sqrt{\varphi^2 + \pi^2 + \rho^2}}.$$

Thus, an *optimal PIR* is one with φ, π and ρ such that

$$\max \frac{\pi + \rho}{\sqrt{2} \cdot \sqrt{\varphi^2 + \pi^2 + \rho^2}}$$

$$\text{subject to } \frac{\varphi\pi}{\rho(1-\pi)} = \frac{\Delta}{M-\Delta}, \quad 0 \leq \rho \leq 1,\ 0 \leq \pi \leq 1.$$

Both methods give, practically, the same global optimum (using MathCAD Plus 8.01 Professional): $\varphi = 0.0000004$, $\rho = 1$, $\pi = 0.999$.

7. Interaction *IR* (I^2R)

Interaction *IR* (I^2R was first defined in [7], [8] based on the concept of interaction in Copenhagen Interpretation (Quantum Mechanics). Given a set $D = \{d_1, d_2, \ldots, d_i, \ldots, d_M\}$, $M > 1$, of objects. A totally bi-directionally connected *Artificial Neural Network* (*ANN*) D-Net $= \langle \aleph, W, L \rangle$ is associated: $\aleph = \{\nu_i : \nu_i$ artificial neuron assigned to object d_i, $i = 1, 2, \ldots, M\}$ denotes a set of artificial neurons, $L : \aleph \times \aleph \to \mathbb{R}_+^{K_{ij}}$, $L(\nu_i, \nu_j)$ denotes connection strengths or weights, $i, j = 1, 2, \ldots, M$, and $W = \{\mathbf{w}_{ij}: i, j = 1, 2, \ldots, M\}$ denotes a set of weights. The following conditions hold: $\mathbf{w}_{ij} \neq \mathbf{0} \Rightarrow \mathbf{w}_{ji} \neq \mathbf{0}$, $\forall i, j, 0 \leq \leq w_{ij}^k \leq 1$, $\forall i, j, k$. The *state* of ν_i is denoted by z_i. An *activation spreading* takes place in D-Net.

Definition 15. *Feeding* a new artificial neuron ν into a D-Net $= \langle \aleph, W, L \rangle$ means obtaining a new D'-Net $= \langle \aleph', W', L' \rangle$ where $\aleph' = \aleph \cup \{\nu\}$, $L' : \aleph' \times \aleph' \to \mathbb{R}_+^{K_{ij}}$.

W' contains more weights than W hence $|W'| > |W|$, $W' \backslash W \neq \emptyset$. One can distinguish three cases: a) W' contains the new weights for the fed ν and all of the old weights unchanged, b) W' contains the new weights for the fed ν and all of the old weights changed, c) W' contains the new weights for the fed ν and both changed and unchanged old weights.

Definition 16. Let $\langle \aleph, W, L \rangle$ be a D-Net and $\langle \aleph', W', L' \rangle$ be a fed D'-Net. The set difference $W' \backslash W$ is *interaction* I between ν and D-Net, $I = W' \backslash W$. When only a) occurs (see above), interaction is called a *pseudo-interaction*: $I = W' \backslash W = \{\mathbf{w}'_{\nu j}, \mathbf{w}'_{j\nu}, \forall j\}$. When b) or c) or both occur (see above), interaction is called a *real interaction*: $I = W' \backslash W = \{\mathbf{w}'_{\nu j}, \mathbf{w}'_{j\nu}, \exists i, j : \mathbf{w}'_{ij} \neq \mathbf{w}_{ij}\}$.

The nonclassical model of *IR* is defined as follows:

Definition 17. Let $\langle \aleph, W, L \rangle$ be a D-Net and $\langle \aleph' = \aleph \cup \{\nu_q\}, W', L' \rangle$ an associated fed D'-Net. *Interaction Information Retrieval* ($IIR = I^2R$)

on D' is a 2-tuple $\langle D', \Re \rangle$ where $\Re(q) = \{d_i : \nu_i \text{ is a winner in an activation spreading started at } \nu_q\}$.

7.1. *CIR* as a special case of I^2R

As a noteworthy property, it is now shown that the classical *IR* is a special case of I^2R.

Theorem 10. *Given an I^2R, an arbitrary fixed $k \geq 1$, a threshold value α_k and ν_q. If* i) $K_{ij} = C = \text{constant} \geq 1$, $\forall i, j$, ii) $L(\nu_i, \nu_i) = \mathbf{1}$, iii) $z_j \leftarrow$ *if* $(w_{qj}^k = \max_p w_{qj}^p) \wedge (w_{qj}^p > \alpha_k)$ *then set* z_j *to 'winner' else* 0, $q \neq j$; iv) *after the first step, stop activation; then I^2R is equivalent to CIR.*

Proof. I^2R with conditions i)–iv) satisfies properties **P1** and **P2** of *CIR*. Condition i) $K_{ij} = C = \text{constant}$, $\forall i, j$ is viewed as the number C of criteria. L is viewed as membership function $\mu : w_{qj}^p = \mu_{\tilde{a}_p}(d_q, d_j)$. Condition ii) $L(\nu_i, \nu_i) = \mathbf{1}$ corresponds to **P1**: $L(\nu_i, \nu_i) = \mathbf{1} \Leftrightarrow w_{ii}^p = 1$, $\forall p \Leftrightarrow \mu_{\tilde{a}_p}(\nu_i, \nu_i) = 1$, $\forall p$. $\Re(q)$ in I^2R is equal to $\Re(q) = \{d_j : \nu_j$ is a winner in an activation started at $\nu_q\}$ which rewrites as $\Re(q) = \{d_j : \nu_j, (w = \max_p w_{pj}^p \wedge (w_{pj}^p > \alpha_k)\} = \{d_j : \nu_j, \mu_{\tilde{a}_k}(\nu_q, \nu_j) = \max_{p=1,\ldots,C} \mu_{\tilde{a}_p}(\nu_q, \nu_j)\} \cap \{d_j : \nu_j, \mu_{\tilde{a}_k}(\nu_q, \nu_q) > \alpha_k\}$. $\Diamond$

It can be said that *CIR* is an I^2R with pseudo-interaction.

7.2. Associative I^2R (AI^2R)

A different type of *IR* is that in which a real interaction takes place.

Definition 18. A *reverberative circle* ς is a sequence $\varsigma = \nu', \ldots, \nu^p, \ldots, \nu^V$ of artificial neurons where: $\nu' = \nu^V$ and ν^p is a winner, i.e. is the most active of all elements succeeding its predecessor ($p-1$ denotes predecessor), i.e. ν^p such that $z_p = \max_j \{z_j : \mathbf{w}_{p-1,j} \neq \mathbf{0}, p-1 \neq j\}$.

Definition 19. An element ν *recalls* a reverberative circle ς if ς is formed due to an activation spreading originating at ν.

Stability in implementation is guaranteed by the following:

Theorem 11. *There exists at least one reverberative circle ς in a D-*Net *recalled by a non-isolated element ν.* $\Diamond$

An I^2R with real interaction is now defined.

Definition 20. Let $\langle \aleph, W, L \rangle$ be a D-Net and $\langle \aleph', W', L' \rangle$ an associated fed D'-Net. *Associative Interaction Information Retrieval* (AI^2R) on D' is a 2-tuple $\langle D, \Re \rangle$ where $\Re(q) = \{d_i : \nu_i \in \varsigma, \varsigma$ recalled by non-isolated $\nu_q\}$.

Specific properties, including implementation, are investigated in [29], [25], [30], [22], [23], [24], [9], [10], [11], [31], [3].

The mathematical structure of retrieval is as follows:

Theorem 12. *Retrieval in AI^2R means defining a matroid.*

Proof. $\langle \aleph', W', L' \rangle$ can be assigned a complete, directed, weighted multigraph G as follows: i) each artificial neuron ν_i is assigned a vertex v_i, ii) there are two oppositely directed edges (opposite arcs), e_{ij} and e_{ji}, between every pair of vertices v_i and v_j $(i \neq j)$ having weights u_{ij} and u_{ji} respectively, where $u_{ij} = \sum_{k=1}^{K_{ij}} w_{ij}^k$ and $u_{ji} = \sum_{k=1}^{K_{ji}} w_{ji}^k$. Any reverberative circle ς corresponds to a circle C in graph G. Let $N = \{\nu_a : a = 1, 2, \ldots, A\}$ denote the artificial neurons traversed before ς is recalled by ν_q. Then N corresponds to a path $P = \{v_a : a = 1, 2, \ldots, A\}$. This means that retrieval defines a connected subgraph H with circles and cutpoints. Hence a block-cutpoint graph T can be assigned to subgraph H which generates its cycle matroid. $\Diamond$

A definition of optimal effectiveness, for an implemented I^2R, can be given as follows. Ideally: $\pi = 1\rho = 1$. Thus:

Definition 21. An (implemented) AI^2R is *optimal* if $\Re(q)$ is a matroid.

8. Special topics

8.1. Boolean and Cluster *IR*.

Let $\langle D, \Re \rangle$ be an I^2R, i.e. AI^2R or *CIR* (*SIR* or *PIR*), and let $(q_k)_{k=1,\ldots,K}$ be a series of objects (-queries), and $Q = \Phi(q_k)$ be a Boolean expression over q_k in a normal form.

Definition 22. *Boolean IR* (*BIR*) over $\langle D, \Re \rangle$ is a 2-tuple $\langle D, \beta \rangle$ where $\beta = \Phi'(\Re_k)\Re_k = \Re(q_k)$. Φ' is a logical counterpart of Φ.

Definition 23. *Cluster IR*: D is a set of cluster representatives, and retrieval any of the previous ones.

8.2. Comparison of *CIR* and *MADM*.

Let $X = \{x_1, x_2 \ldots x_j, \ldots, x_M\}$ be a set of *alternatives*, and $G = \{\tilde{A}_1, \tilde{A}_2 \ldots\ldots \tilde{A}_{i \ldots \tilde{A}_C}\}$ a set of *goals*. Each goal $\tilde{A}_i$ is assigned a *weight* w_i representing its "importance". The attainment of goal $\tilde{A}_i$ by alternative x_j is expressed by the degree of membership $\mu_{\tilde{A}_i}(x_j)$. Decide on an optimal alternative ([40]). The decision $\tilde{N}$ is defined as follows: $\tilde{N} = \oplus \tilde{A}_j^{w_i}$ where $\oplus$ denotes the "confluence operator". The optimal alternative x^* is defined as $\mu_{\tilde{N}}(x^*) = \max_{j=1,\ldots,M} \mu_{\tilde{N}}(x_j)$. The set X corresponds to set D of objects, and the set G of goals to set A of criteria. The criteria do not have weights in *CIR*, $w_i = 1$.

An equivalence between *CIR* and *MADM* (Multi-Attributive Decision Making) can be shown.

Theorem 13. *If* i) *the confluence operator* $\oplus$ *is the fuzzy union,* $w_i = 1$, $\forall i$; ii) i *in property* **P2** *of CIR corresponds to* x^*; iii) $\mu_{\tilde{A}_i}(x_j) \geq \mu_{\tilde{A}_k}(x_j)$, $\forall j, k$ *then CIR and a repeatedly applied MADM, after deleting the already selected alternative, are equivalent.* $\Diamond$

The results in [36], [37] can be consistently build on this theorem.

8.3. Computational complexity in *IR*.

Computational complexity aspects are reflected in the following two theorems.

Theorem 14. *CIR is in the* **P**-*Class.* $\Diamond$

Theorem 15. *AI*^{2}*R is in the* **NP**-*Class.*

Proof. The number $s(M, p)$ of evaluations is $s(M, p) = M \cdot (M - 1 + \sum_{p=1}^{M} \binom{M}{p})$ where M denotes the number of elements and p is the maximum number of most active elements. Because $s(M, p) = M \cdot (M - 1 + \sum_{p=1}^{M} \binom{M}{p}) = M \cdot (M - 1 + 2^M - 1) = O(2^M)$ AI^2R is in the **NP**-Class. $\Diamond$

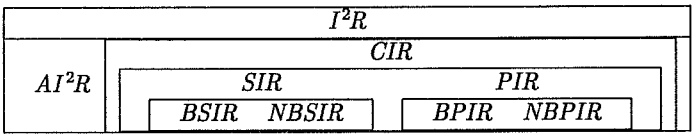
8.4. Comparison between I^2R and Situation *IR*.

Situation *IR* (*SITIR*) is a logic model of *IR* [18] based on Situation Theory [6]. Because *SITIR* aims at an axiomatic approach towards *IR*, a formal comparison between *SITIR* and I^2R can be made. We should note, however, that *SITIR* is not concerned with measuring the uncertainty of the suggested plausible inferences, hence its concept of retrieval is not defined. Thus, the comparison cannot be complete.

The *SITIR* model is defined as follows. Given a set $D = \{d_k : k = 1, 2, \ldots, M\}$ of *documents*. Let $S = \{s_k : k = 1, 2, \ldots, M\}$ be an associated set of *situations*, where situation $s_k = \{i_{kp} : i_{kp} = \langle\langle R_{kp}, a_{k_1}, \ldots, a_{knp}, I_{kp}\rangle\rangle, p = 1, \ldots, p_k\}$, $k = 1, 2, \ldots, M$, is a set of *infons*. Retrieval is not defined in *SITIR* (mathematically nor implementably). A concept of situation aboutness $\rightsquigarrow$ is suggested: $s_k \rightsquigarrow s_i \Leftrightarrow \exists i_r \in s_i[s_k \models i_r]$. Depending on how $\models$ is defined, a Boolean or a *Coordination Level Matching* (*CLM*) model is obtained. In *CLM* this simply reduces to $s_k \rightsquigarrow s_i \Leftrightarrow s_k \cap s_j \neq \emptyset$. As to the Boolean model: we have already seen that this is a *method* of precessing a query rather than a basic model. An element ν_k of I^2R's D-Net is assigned the following situation $s_k = \{i_{kp} : i_{kp} = \langle\langle R_{kp}, \mathbf{w}_{kp}, I_{kp}\rangle\rangle, \forall p \neq k\}$. Situation aboutness $s_k \rightsquigarrow s_i \Leftrightarrow \exists i_r \in s_i[s_k \models i_r]$ is defined as $s_k \rightsquigarrow s_i \Leftrightarrow \mathbf{w}_{jk} \neq \mathbf{0}$.

9. Conclusions, remarks

A unified mathematical foundation and theory for *IR* are created, for the first time. The following is a schematic of their structure (D=definition, T=theorem):



Thus, *IR* becomes a mathematical and hence a natural scientific discipline, too.

SIR's particular structure, the σ-space, is characterized by non-metric topologies in general, and metric induced topologies (in Banach-, Hilbert, Euclidean-spaces) in typical cases, and it should therefore gain mathematical status. The σ-space can also be a link between *SIR* and van Rijsbergen's Information Logic: σ may be viewed as a measure for the minimal extra information principle.

The problem of considering a concept of infinity in *SIR* is to be investigated considering that M and $\Re(q)$ can already be in the millions nowadays (e.g. in Internet searching).

The mathematical formulation of relevance effectiveness (maximal subspace consisting of Cauchy-sequences in *SIR*, recursion fixed point as constrained nonlinear optimisation in *PIR*, and matroid consisting of associated local memories in AI^2R) is also given.

Links between *IR* and modern mathematical disciplines (Functional Analysis, Recursion Theory, Matroid Theory, Complexity Theory, Decision Theory, Fuzzy Sets Theory) are formally created. The manifoldness of these links reflect that of *IR* itself. And yet, *IR* now has its own unified mathematics: language, style, content.

Acknowledgement

This research has partly been supported by Research Grant Pro Renovanda Tud. 98/31. The author wishes to thank the reviewer's precious critical remarks, and Mathematica Pannonica for embracing a new mathematical theory. Special thanks go to many of my colleagues for helpful discussions.

References

[1] BORDOGNA, G. and PASI, G.: A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval: A Model and Its Evaluation. *Journal of the American Society for Information Science* **44**/2 (1993), 70–82.

[2] BUCKLAND, M. and GEY, F.: The relationship Between Recall and Precision, *Journal of the American Society for Information Science* **45**/1 (1994), 12–19.

[3] CARRICK, C. and WATTERS, C.: Automatic Association of News Items, *Information Processing and Management* **33**/5 (1997), 615–632.

[4] CHEN, P.S.: On Inference Rules of Logic-Based Information Retrival Systems, *Information Processing and Management* **30**/1 (1994), 43–59.

[5] da SILVA, W.T. and MILIDIU, R.L.: Belief Function Model for Information Retrieval, *Journal of the American Society for Information Science* **44**/1 (1993), 10–18.

[6] DEVLIN, K.: Logic and Information, Cambridge University Press, Cambridge, Great Britain, (1991).

[7] DOMINICH, S.: The Formulation of the Interaction Information Retrieval Model As A New And Complementary Framework For Information Retrieval Ph.D. Thesis, (in English), Hungarian Academy of Sciences, Budapest, Hungary, 1993.

[8] DOMINICH, S.: Interaction Information Retrieval, *Journal of Documentation* **50**/3 (1994), 197–212.

[9] DOMINICH, S.: The Interaction-Based Information Retrieval Paradigm. In: Kent, A. and Williams, J.G. (eds.) *Encyclopedia of Computer Science and Technology* Vol. 37, Suppl. 22, Marcel Dekker, Inc., New York Basel Hong Kong (1997), 175–192.

[10] DOMINICH, S.: The Interaction-Based Information Retrieval Paradigm. In: Kent, A. (ed.) *Encyclopedia of Library and Information Science* Vol. 59, Suppl. 22, Marcel Dekker, Inc., New York Basel Hong Kong (1997), 218–238.

[11] DOMINICH, S.: An I^2R (Interaction Information Retrieval) Pre-processor for Relevance Feedback, *Technology Letters* **2**/1 (1998), 5–18.

[12] DOMINICH, S.: A Unified Mathematical Definition of Clasical Information Retrieval. To appear in: *Journal of the American Society for Information Science* (2000).

[13] EGGHE, L. and ROUSSEAU, R.: Duality in Information Retrieval and the Hypergeometric Distribution, *Journal of Documentation* **53**/5 (1997), 488–496.

[14] EGGHE, L. and ROUSSEAU, R.: Topological Aspects of Information Retrieval, *Journal of the American Society for Information Science* **49**/13 (1998), 1144–1160.

[15] EVERETT, D.M. and CATER, S.C.: Topology of document retrieval systems, *Journal of the American Society for Information Science* **43**/10 (1992), 658–673.

[16] FUHR, N.: Probabilistic Models in Information Retrieval, *The Computer Journal* **35**/3 (1992), 243–255.

[17] FUHR, N. and ROELLEKE, T.: A probabilistic relational algebra for the integration of information retrieval and database systems, *ACM Transactions On Information Systems* **15**/1 (1997), 32–66.

[18] HUIBERS, T.W.C. and BRUZA, P.D.: Situations, a General Framework for Studying Information Retrieval. In: Leon, R. (ed.) Information Retrieval New Systems and Current Research. Proceedings of the 16th Research Colloquim of the British Computer Society Information Retrieval Specialist Group. Drymen, Taylor Graham (1996), 3-25.

[19] KANG, H.K. and CHOI, K.S.: Two-level Document Ranking Using Mutual Information In Natural Language. *Information Processing and Management* **33**/3 (1997), 289–306.

[20] KRAFT, D.H., BORDOGNA, G. and PASI, G.: Fuzzy Information Retrieval, In: Didier, D. and Prade, H. (eds.) *Handbook of Fuzzy Sets*, Kluwer, 1998.

[21] LEE, J.J. and KANTOR, P.B.: A Study of Probabilistic IR Systems in the Case of Inconsistent Expert Judgement, *Journal of the American Society for Information Science* **42**/3 (1991), 166–172.

[22] LIN, X.: Map Displays for Information Retrieval, *Journal of the American Society for Information Science* **48**/1 (1997), 40–54.

[23] LIU, G.Z.: Semantic Vector Space Model: Implementation and Evaluation, *Journal of the American Society for Information Science* **48**/5 (1997), 395–417.

[24] MOCK, K.J. and VEMURI, V.R.: Information Filtering via Hill Climbing, Wordnet and Index Patterns, *Information Processing and Management* **33**/5 (1997), 633–644.

[25] PEARCE, C. and NICOLAS, C.: TELLTALE: Experiments in a Dynamic Hypertext Environment for Degraded and Multilingual Data, *Journal of the American Society for Information Science* **47**/4 (1996), 263–275.

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

[27] ROBERTSON, S.E., MARON, M.E. and COOPER, W.S.: Probability of relevance: A unification of two competing models for document retrieval, *Information Technology: Research and Develeopment* **1** (1982), 1–21.

[28] SALTON, G. and McGILL, M.: Introduction to Modern Information Retrieval, McGraw Hill, New York, 1983.

[29] SALTON, G., ALLAN, J. and SINGHALL, A.: Automatic Text Decomposition and Structuring, *Information Processing and Management* **32**/2 (1996), 127–138.

[30] SALTON, G., SINGHALL, A., MITRA, M. and BUCKLEY, C.: Automatic Text Structuring and Summarization, *Information Processing and Management* **33**/2 (1997), 193–207.

[31] SHAW, W.M., BURGIU, R. and HOWELL, P.: Performance Standards and Evaluation in Information Retrieval Test Collection: Cluster-based Retrieval Models, *Information Processing and Management* **33**/1 (1997), 1–14.

[32] van RIJSBERGEN, C.J.: Information Retrieval. Butterworth, London, 1979.

[33] van RIJSBERGEN, C.J.: Probabilistic Retrieval Revisited, *The Computer Journal* **35**/3 (1992), 291-298.

[34] van RIJSBERGEN, C.J. and LALMAS, M.: An Information Calculus for Information Retrieval, *Journal of the American Society for Information Science* **47**/5 (1996), 385–398.

[35] WONG, S.K.M. and YAO, Y.Y.: A Generalized Binary Probabilistic Independence Model, *Journal of the American Society for Information Science* **41** (1990), 324–329.

[36] WONG, S.K.M. and YAO, Y.Y.: A Probabilistic Inference Model for Information Retrieval based on Axiomatic Decision Theory, *Information Systems* **16** (1991), 301–321.

[37] WONG, S.K.M., BOLLMANN, P. and YAO, Y.Y.: Information Retrieval based on Axiomatic Decision Theory, *General Systems* **19** (1991), 101–117.

[38] WONG, S.K.M. and YAO, Y.Y.: A Probabilistic Method for Computing Term-by-Term Relationships, *Journal of the American Society for Information Science* **44**/8 (1993).

[39] YU, C.T., MENG, W. and PARK, S.: A Framework for Effective Retrieval, *ACM Transactions on Database Systems* **14** (1989), 147–167.

[40] ZIMMERMAN, H.J.: Fuzzy Set Theory - and Its Applications, Kluwer Academic Publishers, Boston, Dordrecht, London, 1996