Formal Foundation of Information Retrieval

Sandor Dominich

Department of Computing and Information Technology, Buckinghamshire Chilterns University College, High Wycombe, United Kingdom, Queen Alexandra Road, HP11 2JZ, E-mail: sdomin01@bcuc.ac.uk; Department of Computer Science, University of Veszprem, Hungary, 8200 Veszprem, Egyetem u. 10, E-mail: dominich@dcs.vein.hu

In any information retrieval system, retrieval is based on some formal model. Apart from a few – non-classical – models (which are relatively newer), every other model ultimately relies on two basic models: vector space, and probabilistic (two early and classical models). Hence the vector space and probabilistic models are of a fundamental importance, and thus a unified formal definition for them would allow for working out a unified, coherent and consistent formal framework (mathematical theory), as a foundation, for IR. The paper shows that such foundations can be elaborated.

1. Introduction

Information Retrieval (IR) has accumulated a huge amount of experimental and theoretical results, and its importance has rapidly grown especially with the advent of the World Wide Web (WWW) and Internet.

IR is an interdisciplinary field, and thus an attempt to elaborate formal foundations for it should take into account a manifold of viewpoints, techniques and parameters coming from a variety of fields.

The goal of the present paper is much more restricted: it aims at giving, in an axiomatic style, a formal (mathematical) foundation for the classical (vector space, probabilistic) models of IR, i.e. a unified definition from which all known mathematical properties can be formally derived.

The paper consists of two parts with the same content but different style. In Part I, the formal foundations are presented informally, whilst in Part II the formal foundations are presented using a mathematical formalism.

PART I.

2. Traditional Definitions of Classical Models

2.1 Vector Space Model

Both the *query* and the *document* are represented as series of *weights* (vectors). Every weight represents a degree to which an *index term* (e.g. keyword, identifier) characterizes (pertains to) a – real – document (or query). A *similarity measure*, as a computable numeric value, is defined – using the query- and document-vector – to express a *likeness* (closeness) between a query and a document. The similarity measure typically has the following three basic properties: (i) It usually takes on values between 0 and 1; (ii) Its value does not depend on the order in which the query and the document are compared (when computing the similarity); (iii) It is equal to 1 when the query- and document-vectors are equal. Those documents are said to be *retrieved* in response to a query for which the similarity measure exceeds a *threshold* value.

2.2 Probabilistic Model

Given *documents*, a *query*, and a *cut-off* numeric value. Conditional *probabilities* that a document is *relevant* and *irrelevant* to the query are calculated. The documents with probabilities of relevance *at least* that of irrelevance are *ranked* in decreasing order of their relevance. Those documents are said to be *retrieved* whose probabilities of relevance in the ranked list *exceed* the cut-off value.

3. Unified Formal Definition of Classical Models

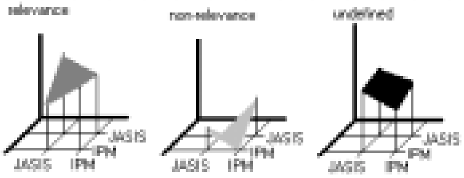
3.1 Classical Retrieval

Given a finite set of elements called *identifiers* (e.g. index terms, keywords, descriptors), a finite set of elements called *documents* (e.g. text, image, pieces of sound) where a document (i) represents a collection (cluster) of some other objects (as a particular case: a document can, of course, represent itself), and (ii) is modeled as a series of numbers (weights) meaning degrees to which the identifiers pertain to that document. Given further a finite set of elements called *criteria* (e.g. relevant, irrelevant, undefined) according to which every two documents are compared to each other, and the result of this comparison is a numeric value (score) assigned to that pair of documents. A concept of *classical retrieval* is defined as a process (e.g. mapping, function, procedure) whereby, given a criterion, those documents are associated (found, returned, retrieved) to a query (which is formally a document) for which the scores are greatest.

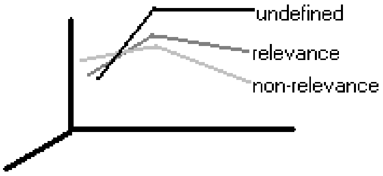
Example. Let the set of identifiers be T = {computer, information, library} and the set of documents be O = {JASIS, IPM}, thus D can be modelled as shown in the table.

	computer	information	library
JASIS	1	1	0.5
IPM	0.8	1	0.5

Each criterion can be represented by a surface. The two documents, JASIS and IPM, are represented on the horizontal axes. The values corresponding to a criterion are measured on the vertical axis, yielding a surface.



Taking a query q is equivalent to cutting the corresponding surface with a vertical plane (which corresponds to q). The result will be a curve in space. There are as many curves as criteria. If all curves are represented in one system of coordinates one gets a visual impression of the formal definition of classical retrieval.



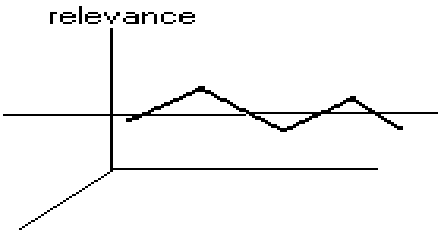
The concept of classical retrieval may be viewed as a formal superstructure with an underlying parameter: the number of criteria. Depending on the values chosen for this parameter, one can get the vector space or the probabilistic model.

Note

We should note, however, that, from a purely formal point of view, there is another parameter, too, and this is whether the order in which the query and document are considered for evaluation (specifically, the computation of similarity or conditional probability) matters or not; this is mathematically expressed by a property called commutativity. Because, on the one hand, trying to visualise the influence of this parameter in our example seems very difficult, and on the other hand not visualising it does not hinder the understanding, we will only be considering it in the formal section.

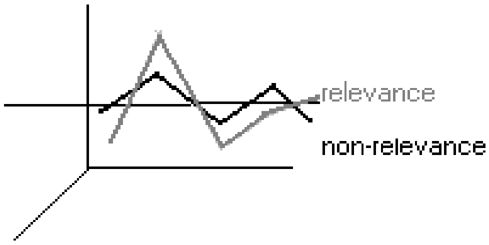
3.2 Vector Space Model

If there is just one criterion (which may be called e.g. relevance), and the evaluation order of query and document does not matter, one gets the vector space model. Retrieval means that part of this curve, which is above the straight line corresponding to a threshold value. This view is equivalent to the traditional definition.



3.3 Probabilistic Model

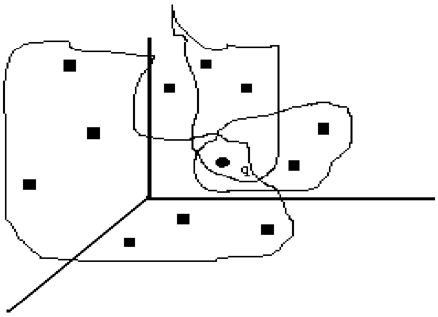
If there are two criteria, called e.g. relevance and non-relevance, and the evaluation order of query and document does matter (namely: document to query), one gets the probabilistic model. Visually, this corresponds to two curves in space. Retrieval is defined with respect to one fixed criterion (of the two) as being those parts of the corresponding curve that are above the other curve and exceed a cut-off value (which again can be represented by a straight line). This view is equivalent to the traditional definition.



4. Mathematical Structures

4.1 Vector Space Model

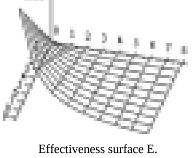
Let *d* denote a threshold value, and *V* denote the corresponding set of retrieved documents (for a given query *q*). We may say in other words, that the documents of *V* constitute a *neighborhood* of the query. By taking different values for *d*, e.g. d_1, d_2, ..., different sets of neighborhoods are obtained, e.g. V_1, V_2, .. Thus, retrievals may be conceived as defining a series of such neighborhoods of the query in the space of documents.



4.2 Probabilistic Model

Retrieval is driven by a sequence of repeatedly applied *relevance feedback*, and so there is a sequence of corresponding triplets (*precision, recall, fallout*). The consecutive application of relevance feedback is performed in a recursive manner: the set of documents retrieved at one relevance feedback enters the next relevance feedback process. Thus, the consecutive triplets (precision, recall, fallout) correspond to points on a surface, and they define a 'walk' towards an optimal point.

The vertical axis corresponds to fallout, the axis to its right to precision, and the third axis corresponds to recall. Consecutive values of the 3-tuple (precision, recall, fallout) define a surface in this space, which may thus be termed as an effectiveness surface. The aim of retrieval is to reach as low a point on the effectiveness surface as possible in as stable a way as possible. It can be seen that fallout increases when precision is low and recall is high. The lower on the surface, the higher precision and/or recall.



Effectiveness surface E.

E

4.2 Boolean Model

In the *Boolean model*, the query is a Boolean-type expression of terms. Retrieval consists of two steps: (i) first, for every query term, those documents are retrieved which contain (or not, depending on the operator used) that term, and then (ii) the thus retrieved sets of documents are subjected to set theoretic operations (corresponding to the logical operators in the query). Step (i) is a document selecting condition, and is a very special type of similarity (which, for example, is equal to 1 if cosine≠0, and zero otherwise). Thus, the mathematical structure characterizing retrievals in the Boolean model depend on the vector space model used for the document selecting condition. (Because of this, the Boolean model is not a true elementary model of IR.)

In the – really – *weighted Boolean model*, query terms are associated weights expressing their relative importance. Different *evaluation measures* have been suggested for the document selecting condition, of which none is a classical measure (similarity, probability), and hence the weighted Boolean model seems to be a true – non-classical – model of IR.

PART II.

5. Traditional Definitions of Classical Models

Definition of Vector Space IR (VIR). Let D be a set – representations – of *documents.* A function $\sigma: D \times D \to [0; 1]$ is called a *similarity* if the following properties hold: (i) $0 \le \sigma(a, b) \le 1$, $\forall a, b \in D$ (*normalization*); (ii) $\sigma(a, b) = \sigma(b, a)$, $\forall a, b \in D$ (*commutativity*); (iii) $a = b \Rightarrow \sigma(a, b) = 1$, $a, b \in D$ (*reflexivity*). Let $q \in D$ be a *query*, and $\tau \in \mathbf{R}$ be a real *threshold* value. The set $\Re(q)$ of *retrieved* documents in response to query q is defined as follows: $\Re(q) = \{d \in D | \sigma(d, q) > \tau\}$.[e.g., Bollmann-Sdorra and Raghavan; Buckland; Caid, Dimais and Galiant; Egghe and Rousseau; Everett and Cater; Kang and Choi; Kowalski; Luhn; Raghavan and Wong; Raghavan , Wong and Ziarko; Salton; Salton and McGill; van Rijsbergen]

Definition of Probabilistic IR (PIR). Let D be a set of *documents*, $q \in D$ a *query*, $\alpha \in \mathbf{R}$ a real *cut-off value*, and $P(R|(q, d))$ and $P(I|(q, d))$ the probability that document d is relevant (R) and irrelevant (I), respectively, to query q. It is assumed that $P(R|(d, d)) = 1$, $P(I|(d, d)) = 0$. The *retrieved* documents in response to query q belong to the set $\Re(q)$ defined as follows: $\Re(q) = \{d | P(R|(q, d)) \ge P(I|(q, d)), P(R|(q, d)) > \alpha\}$. [e.g., Callan, Croft and Harding; Cooper and Maron; Croft; Harding and Weir; Fuhr, Huang and Robertson; Maron and Kuhns; Robertson and Sparck-Jones; Robertson, Maron and Cooper; van Rijsbergen; Wong and Yao; Wong,, Butz and Xiang; Yu, Meng and Park]

6. Unified Formal Definition of Classical Models

Given: a finite set $T = \{t_1, t_2, ..., t_k, ..., t_N\}$ of elements called *identifiers*, $N \ge 1$; a finite set $O = \{o_1, o_2, ..., o_u, .., o_U\}$ of elements called *objects*, $U \ge 2$; a finite family $(D_j)_{j \in J=\{1, 2, ...,M\}}$ of object *clusters*, $D_j \in \wp(O)$, $M \ge 2$; a set $D = \{\tilde{o}_j | j \in J\}$ of *documents* where the normalized fuzzy set $\tilde{o}_j = \{(t_k, \mu_{\tilde{o}_j}(t_k)) | t_k \in T, k = 1, 2, ..., N\}$, $j = 1, 2, ..., M$, $\mu_{\tilde{o}_j}: T \to S \subseteq [0; 1] \subset \mathbf{R}$, is a cluster *representative* of object cluster D_j; a finite set $A = \{\tilde{a}_1, \tilde{a}_2, ..., \tilde{a}_i, ..., \tilde{a}_C\}$ of *criteria*, $C \ge 1$, where $\tilde{a}_i = \{((q, \tilde{o}_j), \mu_{\tilde{a}_i}(q, \tilde{o}_j)) | \tilde{o}_j \in D, j = 1, 2, ..., M\}$, $i = 1, 2, .., C$, is a normalized fuzzy relation, $\mu_{\tilde{a}_i}: D \times D \to [0, 1] \subset \mathbf{R}$, $q \in D$ arbitrary fixed; an α_i-cut $a_{\alpha_i} = \{\tilde{o} \in D | \mu_{\tilde{a}_i}(q, \tilde{o}) > \alpha_i\}$, $i = 1, 2, ..., C$, of criterion $\tilde{a}_i$, $0 \le \alpha_i < +\infty$; and a mapping $\mathfrak{R}: D \to \wp(D)$ called *retrieval*. [e.g., Kraft, Bordogna and Pasi; Miyamoto; Wong and Yao; Wong, Bollmann and Yao; Dominich]

Definition 6.1. A 2-tuple $\langle D, \Re \rangle$ with the following properties: (P1) $q = \tilde{o} \Rightarrow \mu_{\tilde{a}_i}(q, \tilde{o}) = 1$ (reflexivity), $q, \tilde{o} \in D$; (P2) $\mathfrak{R}(q) = \{\tilde{o} | \mu_{\tilde{a}_i}(q, \tilde{o}) = \max_{k=1,...,C} \mu_{\tilde{a}_k}(q, \tilde{o})\} \cap a_{\alpha_i}$, is called a *Classical Information Retrieval* (*CIR*); $\tilde{a}_i$ arbitrary fixed. [Dominich]

7. Vector Space Model

The vector space model is defined as a special case of *CIR* as follows.

Definition 7.1. *The vector space model* (VSM) is a *CIR* $\langle D, \Re \rangle$ obeying the following conditions: (S1) $C = 1$; (S2) $\mu_{\tilde{a}_i}$ is commutative, i.e. $\mu_{\tilde{a}_i}(q, \tilde{o}) = \mu_{\tilde{a}_i}(\tilde{o}, q)$, $\forall \tilde{o}, q \in D$. [Dominich]

It can be shown that the vector model as defined in Def. 7.1 is equivalent to the traditional definition (VIR) of the classical vector space model, and thus the latter is a special case of *CIR*.

THEOREM 7.1. VIR and VSM are equivalent♦ [Dominich]

The vector space model has two particular cases: binary and non-binary. These can easily be defined as being special cases of VSM as follows.

Definition 7.2. The *binary vector space model* is a VSM $\langle D, \Re\rangle$ with $S = \{0, 1\}$. The *non-binary vector space model* is a VSM $\langle D, \Re\rangle$ with $S = [0; 1]$.

It is convenient (and easy to see it) to re-define the vector space model by specifying retrieval using a similarity as follows.

Definition 7.3. Let D be a set of documents, and σ a similarity on D. A *similarity space* (or σ-*space*) on D is a 2-tuple $\Sigma = \langle D, \sigma\rangle$, and a *vector space* (or *similarity*) *model* on D is a σ-space with $\Re(d)=\{x \in D|\sigma(d, x) > \tau, \tau \in \mathbf{R}\}$.

The following – naturally arising – connections to metric and topological spaces can be shown. [e.g., Bollmann-Sdorra and Raghavan; Egghe and Rousseau; Everett and Cater; Dominich]

THEOREM 7.2. Let $\langle E, \mu\rangle$ be a pseudometric space (μ bounded by 1). Then $\langle E, 1 - \mu\rangle$ is a σ-space. ♦

Theorem 7.2 makes the following definition of an induced σ-space possible.

Definition 7.4. Let $\langle E, \mu\rangle$ be a pseudometric space. Then $\Sigma = \langle E, 1 - \mu\rangle$ is the σ-space *induced* on E by pseudometric μ.

Then, it can be shown that:

THEOREM 7.3. Let $\langle E, \mu\rangle$ be a pseudometric space. The induced topological space is a *VSM* on E. ♦

THEOREM 7.4. Let $\langle E, \mu\rangle$ be a pseudometric space. Then the relation ~ defined as $x \sim y \Leftrightarrow \mu(x, y) = 0$, $\forall x, y \in E$, is an equivalence relation on E. ♦

THEOREM 7.5 Given a pseudometric space $\langle E, \mu\rangle$, and the relation $x \sim y \Leftrightarrow \mu(x, y) = 0$, $\forall x, y \in E$. Then the space $\langle E^*, \mu^* = \mu\rangle$ is a metric space with $\mu^*(A, B) = \mu(x, y)$, $A, B \in E^*$, $x \in A, y \in B$. ♦

THEOREM 7.6. The Hausdorff space induced by metric μ^* is a *VSM*. ♦

THEOREM 7.7. 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 *VSM* on E are equivalent. ♦

Specific theorems and properties, as well as applications can be unfolded, formally, from the above basic theorems for both the binary and non-binary vector space models.

8. Probabilistic Model

The probabilistic model is defined as another special case of *CIR* as follows.

Definition 8.1. The *probabilistic model* is a *CIR* $\langle D, \Re\rangle$ satisfying the following condition: $C = 2$. [Dominich]

The following theorem is needed first (a connection of direct proportionality between membership μ and probability P).

THEOREM 8.1. Let P be a probability measure defined in a σ-algebra of universe T, and $p_{kj}^{(i)} = P(X_k = \mu_{\tilde{o}_j}(t_k))$, $i = 1, 2$. If (a) $\mu_{\tilde{a}_i}(q, \tilde{o}_j) = \sum_{k=1}^{N} \log (p_{kj}^{(1)}/p_{kj}^{(2)})$, $i = 1,2$; (b) identifier occurrences are independent; (c) the two criteria are disjoint; Then (1) $\mu_{\tilde{a}_1}(\mathrm{q}, \tilde{o}_\mathrm{j}) \geq \mu_{\tilde{a}_1}(\mathrm{q}, \tilde{o}_\mathrm{s}) \Leftrightarrow P(\tilde{a}_1|\tilde{o}_\mathrm{j}) \geq P(\tilde{a}_1|\tilde{o}_\mathrm{s})$; (2) $\mu_{\tilde{a}_1}(q, \tilde{o}_\mathrm{j}) \geq \mu_{\tilde{a}_2}(q, \tilde{o}_\mathrm{j}) \Leftrightarrow P(\tilde{a}_1|\tilde{o}_\mathrm{j}) \geq P(\tilde{a}_2|\tilde{o}_\mathrm{j})$. ♦ [Dominich]

As a special case of Theorem 8.1 the following holds:

THEOREM 8.2. Given documents D_i, $i =1, ..., M$, as frequency vectors $D_i = \{d_{ik}|k = 1, 2, ..., N\}$ and a dot product similarity function $f(X, Y) = \sum_{k=1}^{N} x_k y_k$. Let $D'_i = \{d'_{ik}|k = 1, 2, ..., N\}$ be a set of new frequency vectors with $d'_{ik} = log(P(d_{ik}|R)/P(d_{ik}|I))$ where $P(d_{ik}|.)$ is the probability that a relevant/irrelevant document has d_{ik} occurrences of the kth term. Then $P(D_i|R) > P(D_j|I) \Leftrightarrow f(Q, D_i) > f(Q, D_j)$ where Q denotes a query with all weights equal to 1 for terms present and 0 otherwise. ♦ [Yu, Meng and Park]

It can be shown that the probabilistic model as defined in Def. 8.1 is equivalent to the traditional definition PIR (in the typical case of optimal retrieval).

THEOREM 8.3. PIR and Def. 8.1 are equivalent. ♦ [Dominich]

In the probabilistic model retrieval is driven by relevance feedback. It can be shown that:

THEOREM 8.4. Repeatedly applying relevance feedback in the probabilistic model yields a Diophantine set (of sets of retrieved documents) with a fixed point (corresponding to an optimal solution). ♦ [Dominich]

9. Boolean Model

Given a query Q as a logical expression $Q = L(\theta_k) = \wedge_{j\in J} (\vee_{k\in K} (\theta_k, w_k))$, $\theta_k \in \{t_k, \neg t_k\}$, where $w_k \in [0; 1]$ is a weight associated to and expressing a relative importance of t_k. [e.g., Bordogna and Pasi; Bordogna, Carrara and Pasi; Buell; Kraft and Buell; Kraft et al.; Kraft, Bordogna and Pasi; Kraft; Yager; Salton and McGill; Salton; Turtle and Croft; van Rijsbergen]

DEFINITION 9.1. A *Boolean model* associated to $\langle D, \Re\rangle$ is a pair $\langle D, \beta\rangle$, $\beta = \lambda(S_k)$, where λ denotes a set theoretic counterpart of L, and $S_k = \Re(\theta_k)$.[Dominich]

Depending on how L, λ and $\Re$ are specified, one gets different particular Boolean models: (i) traditional Boolean model for $w_k \in \{0, 1\}$; (ii) weighted Boolean model when $w_k \in [0; 1]$.

In the traditional Boolean model, $S_k = \Re(\theta_k) = \{\text{doc}|\theta_k \in \text{doc}\}$, and this document selecting condition can be conceived as a special similarity (for example, $\sigma = 1$ if cosine $\neq 0$, and zero otherwise). Thus, S_k is a result of a vector space retrieval. Hence, it can be shown that:

THEOREM 9.1. Taking all possible Boolean retrievals over D is equivalent to a similarity space. ♦[Egghe and Rousseau]

THEOREM 9.2. The set of all possible Boolean retrievals, based on elementary queries, using threshold, is equivalent to the retrieval topology; using close matches, is equivalent to the Euclidean topology. ♦ [Egghe and Rousseau]

Note

We may, hence, say that the traditional Boolean model is not a true elementary classical model of IR: it relies on (it needs) a vector space retrieval (for the document selecting condition; S_k). In principle, any other model can be used to get S_k. The classical Boolean model may also be defined as a very special *fusion*, as follows.

Given $CIR_i = \langle D_i, \cup \{\text{query}\}, \Re\rangle$, $i = 1, 2, ..., n$. *Data fusion* is a $CIR = \langle D \subseteq \cup_{i=1,2,...,n} \Re_i(\text{query}), \Re\rangle$. A *Boolean model* is a CIR with $\Re = \beta$.

In the weighted Boolean model, $S_k = \Re(\theta_k) = \{\text{doc}_j|g(w_k, \mu_{jk}) \geq \varepsilon\}$where g is an evaluation measure for the document selecting condition. The *g*s proposed so far suggest that the – really – weighted Boolean model is indeed a distinct – non-classical – model of IR.

10. Conclusions

A unified formal definition for the classical – vector space, probabilistic – models was given in an axiomatic style. All known mathematical results and properties can be formally derived from/build upon it. There only are two true elementary classical IR models: vector space and probabilistic; from a formal point of view, the difference between them is commutativity and the number of criteria. The traditional Boolean model is not a true elementary model of IR (as we perhaps thought), whilst the weighted Boolean model is a true – non-classical – model. The mathematical structure and properties of the weighted Boolean model are to be researched. Due to this foundation, IR can be conceived as a mathematical discipline (too), researchable by rigorous means, methodically teachable to and accessible by a more diverse scientific audience. The concept of a similarity space (σ-space) should perhaps gain a – mathematical – status within it.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval*. Addison Wesley Longman Publishing Co. Inc.

Bollmann-Sdorra, P. and Raghavan, V.V. (1993) On the Delusiveness of Adopting a Common Space for Modelling Information Retrieval Objects: Are Queries Documents?. *Journal of the American Society for Information Science*, 44(10): 579-587.

Bordogna, G. and Pasi, G. (1993) A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval: A Model and Its Evaluation. *Journal of the American Society for Information Science*, 44(2): 70-82.

Bordogna, G. and Pasi, G. (1995) Linguistic aggregation operators in fuzzy information retrieval. *International Journal of Intelligent Systems*. 10(2): 233-248.

Buckland, M.K. (1997) What is a document? *Journal of the American Society for Information Science,* 48(9): 804-809.

Caid, W.R., Dimais, S.T. and Galiant, S.I. (1995) Learned Vector Space Models for Document Retrieval. *Information Processing and Management*. 31(3): 419-429.

Callan, J.P., Croft, W.-B., Harding, S.-M. (1992) The INQUERY retrieval system. In *Proceedings of the 3rd DEXA*. 78-83.

Cooper , W.S. and Maron, M.E. (1978) Foundation of probabilistic and utility-theoretic indexing . *Journal of the Association for Computing Machinery,* 25, 67-80.

Croft, W.B. (1993) Knowledge-Based and Statistical Approaches to Text Retrieval. *IEEE Expert*, April, 8-12.

Dominich, S. (2000). *Mathematical Foundations of Information Retrieval*. Kluwer Academic Publishers.

Dominich, S. (2000). A Unified Mathematical Definition of Classical Information Retrieval. *Journal of the American Society for Information Science*, **51**(7): 614-625.

Dominich, S. (2000). Foundation of Information Retrieval. *Mathematica Pannonica*, **11**(1): 137-153.

Egghe, L. and Rousseau, R. (1998) Topological Aspects of Information Retrieval. *Journal of the American Society for Information Science*, 49(13): 1144-1160.

Everett, D.M. and Cater, S.C. (1992) Topology of document retrieval systems. *Journal of the American Society for Information Science*, 43(10): 658-673.

Fuhr, N. (1992) Probabilistic Models in Information Retrieval. *The Computer Journal*, 35(3), 243-255.

Fuhr, N. and Rolleke, T. (1997) A probabilistic relational algebra for the integration of information retrieval and database systems. *ACM Transactions On Information Systems*, 15(1): 32-66.

Gordon, M.D. and Kochen, M. (1989) Recall-Precision trade-off: A derivation. *Journal of the American Society for Information Science*, 40: 145-151.

Gordon, M.D. (1990) Evaluating the Effectiveness of Information Retrieval Systems Using Simulated Queries. *Journal of the American Society for Information Science*, 41(5): 313-323.

Kang, H.K. and Choi, K.S. (1997) Two-level Document Ranking Using Mutual Information In Natural Language. *Information Processing and Management*. 33(3): 289-306.

Kolmogoroff, A. (1950) *Foundation of Probability*. New York.

Korfhage, R.R. (1997) *Information Storage and Retrieval*. Wiley, New York.

Kowalski, G. (1997) *Information Retrieval Systems: Theory and implementation*. Kluwer Academic Publishers, Boston, Ma.

Kraft, D.H. and Boyce, B.R. (1995). Approaches to Intelligent Information Retrieval. *In*: Petry, F.E. and Delcambre, M.L. (eds.) *Advances in Databases and Artificial Intelligence, vol. 1: Intelligent Database Technology: Approaches and Applications*. JAI Press, Greenwich, CT, 243-261.

Kraft, D.H., Bordogna, P. and Pasi, G. (1995) An extended fuzzy linguistic approach to generalize Boolean information retrieval. *Journal of Information Sciences, Applications.* 2(3): 119-134.

Kraft, D.H., Bordogna, P. and Pasi, G. (1998). Fuzzy Set Techniques in Information Retrieval. *In*: Didier, D. and Prade, H. (eds.) *Handbook of Fuzzy Sets and Possibility Theory. Approximate Reasoning and Fuzzy Infomation Systems*. Kluwer Academic Publishers, AA Dordrecht, The Netherlands, Chp. 8.

Luhn, H.P. (1959). Keyword-in-Context Index for Technical Literature (KWIC Index). In Hays, D.-D. (ed.). *Readings in Automatic Language Processing*. American Elsevier Publishing Company, Inc., 1966, pp. 159-167.

Maron, M.E. and Kuhns, J.-L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7: 219-244.

Miyamoto, S. (1998) Application of Rough Sets to Information Retrieval. *Journal of the American Society of Information Science*. 49(13): 195-205.

Miyamoto, S. (1990) *Fuzzy Sets in information retrieval and cluster analysis*. Kluwer, Dordrecht.

Mizzaro, S. (1997) Relevance: The Whole History. *Journal of the American Society of Information Science*. 48(9): 810-832.

Nie, J.Y. (1992) Towards a probabilistic modal logic for semantic-based information retrieval. *ACM SIGIR 15th International Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, ACM Press, New York, 140-151.

Nie, J.Y., Brisebois, M. and Lepage, F. (1995) Information retrieval as counterfactual. *The Computer Journal*, 38(8): 643-657.

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

Radecki, T. (1976) Mathematical model of information retrieval system based on the concept of fuzzy thesaurus. *Information Processing and Management*, 12: 131-318.

Raghavan, V.V. and Wong, S.K.M. (1986) A Critical Analysis of Vector Space Model for Information Retrieval. *Journal of the American Society for Information Science*, 37: 279-287.

Raghavan, V., Wong, S.K.M. and Ziarko, W. (1990) Vector Space Model in Information Retrieval. *Encyclopedia of Computer Science and Technology*, 22: 423-446.

Robertson, S.E. and Spark Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society of Information Science*. 27: 129-146.

Robertson, S.E., Maron, M.-E. and Cooper, W.S. (1982). Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: research and Development*, 1, 1-21.

Rocchio (1971) Relevance Feedback in Information Retrieval. In Salton, G. (ed.) *The SMART Storage and Retrieval System.* Englewood Cliffs, Prentice Hall, 313-323.

Salton, G. (1965) Automatic Phrase Matching. In: In Hays, D.D. (ed.). *Readings in Automatic Language Processing*. American Elsevier Publishing Company, Inc., 1966, pp. 169-189.

Salton, G. (1968). *Automatic Information Organisation and Retrieval*. McGraw Hill, New York.

Salton, G., McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw Hill, New York.

Salton, G. (1971). *The SMART Retrieval System - Experiment in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, New Jersey.

Smyth, M.B. (1992) Topology. In: Abramsky, S. and Gabbay, D.M. and Maibaum, T.S.E. (eds.) *Handbook of Logic in Computer Science.* Vol. 1, Oxford Scince Publications, Clarenden Press.

Tarski, A. (1956) Fundamental concepts of the methodology of the deductive sciences. In *Logic, Semantics, Metamathematics. Papers from 1923 to 1938.* Clarendon Press, Oxford, 60-109.

van Rijsbergen, C.J. (1979). *Information Retrieval*. Butterworth, London.

van Rijsbergen, C.J. (1986a) A New Theoretical Framework for Information Retrieval. *SIGIR Forum*, 21(1-2): 23-29.

van Rijsbergen, C.J. (1992) Probabilistic Retrieval Revisited. *The Computer Journal*, 35(3): 291-298.

Willard, S. (1970) *General Topology*. Reading, MA, Addison-Wesley.

Wong, S.K.M. and Yao, Y.Y. (1990). A Generalized Binary Probabilistic Independence Model. *Journal of the American Society for Information Science*, 41: 324-329.

Wong, S.K.M. and Yao, Y.Y. (1991). A Probabilistic Inference Model for Information Retrieval based on Axiomatic Decision

Theory. *Information Systems*, 16: 301-321.

Wong, S.K.M., Bollmann, P. and Yao, Y.Y. (1991). Information Retrieval based on Axiomatic Decision Theory. *General Systems*, 19: 101-117.

Wong, S.K.M. and Yao, Y.Y. (1993) A Probabilistic Method for Computing Term-by-Term Relationships. *Journal of the American Society for Information Science*, 44(8):

Wong, S.K.M. and Yao, Y.Y. (1995). On modelling information retrieval with probabilistic inference. *ACM Transactions On Information Systems*, 13(1): 38-68

Wong, S.K.M., Butz, C.J. and Xiang, Y. (1995) A Method for Implementing a Probabilistic Model as a Relational Database. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, Montreal, 156-164.

Yu, C.T., Meng, W. and Park, S. (1989). A Framework for Effective Retrieval. *ACM Transactions on Database Systems*, 14: 147-167.

Zadeh, L.A. (1984) Fuzzy sets and Applications. In Yager, O. et al. (eds.) *Selected Papers by L.A Zadeh*. John Wiley and Sons, Inc., 29-44.

Zadeh. L.A. (1989) Knowledge representation in Fuzzy Logic. *IEEE Transactions on Knowledge and Data Engineering*. 1, 89-100.

Zimmerman, H.J. (1996). *Fuzzy Set Theory - and Its Applications*. 3rd edition. Kluwer Academic Publishers, Boston/Dordrecht/London.