An Outline of a General Model for Information Retrieval Systems

Jianyun Nie
Laboratoire Génie Informatique - IMAG
BP. 53X - 38401 Grenoble Cedex
France
e.mail : nie@imag.imag.fr

This paper is a contribution to the construction of a general model for information retrieval. As in the paper of Van Rijsbergen ([RIJ86]), the implicit base in all information retrieval systems is considered as a logical implication. The measure of correspondence between a document and a query is transformed into the estimation of the strength (or certainty) of logical implication. The modal logics will show its suitability for representing the behavior of information retrieval systems. In existing Information Retrieval models, several aspects are often mixed. A part of this paper is contributed to separate these aspects to give a clearer view of information retrieval systems. This general model is also compared with some existing models to show its generality.

I. INTRODUCTION

In Information Retrieval, some well-known models have been constructed, such as the vectorial model, the boolean model, the probabilistic model and the fuzzy model etc. These models have found many uses in real or experimental information retrieval systems. Among the best-known, we can find SMART ([SAL71]) as a vectorial model, TEXTO ([CHE82]) and MEDLARS ([NLM79]) as boolean models, and so on. These models, however, are all strongly related to a specific data representation. The corresponding retrieval model can hardly be applied to a system having another data representation. There is thus a lack of generality in these systems. Consequently, we have a number of different models which cannot be easily related. This situation is a handicap considering the improvement of Information Retrieval systems. Being aware of this, many people have worked on more general retrieval models over the last years. These studies are generally carried out in two ways: extending an existing model to become more general, or constructing a new general model. The extended boolean model ([WAL79]), for instance, can be considered in the first way. Among the studies related to the second way, we can mention the model of Dabrowski [DAB75], and recently a non-classical logic model ([RIJ86]) proposed by Van Rijsbergen.

In this paper we attempt to design a new retrieval model which is based on Rijsbergen's model. We share most of Rijsbergen's ideas, the most important being that in Information Retrieval, the implicit background is logics, even if it is not yet well formalized. This paper is a contribution to the formalization of the basic function of

Information Retrieval Systems, the correspondence between queries and documents. A general model based on modal logics is proposed and compared with some existing models.

II. RIJSBERGEN'S MODEL

In this model, the basic operation is considered as the comparison of a document with a query. A document is considered as a set of sentences which are interpreted into a predefined semantics. So is the query, being usually a single sentence. For a document to be a "right" one for answering the query, it must "imply" the query. In information retrieval, this implication is always "plausible" rather than "strict". So a measurement of implication strength (or uncertainty measurement) has to be associated with it. To estimate if a document corresponds well to a query, one has to measure the implication certainty (or strength of implication), that is $P(D \rightarrow Q)$. Here D represents a document, Q a query and P a function on strength measurement. The symbol "→" does not signify the same thing as the "material implication" in classical logics. Van Rijsbergen proposes a conditional logical evaluation for this implication as expressed in the following fomulae.

$$P(D \rightarrow Q) = P(D \mid Q) \qquad (1)$$

When D and Q are represented by an independent set of index terms, and n represents the cardinality of a set of terms, the formulae 1 may be expressed as:

$$P(D \rightarrow Q) = \frac{n(D \cap Q)}{n(Q)}$$

which expresses that the strength (certainty) of implication is measured as the proportion of the query terms found in the document.

III. SOME PROPOSALS ABOUT AN INFORMATION RETRIEVAL MODEL

Suppose that we have a set of documents. We all know that when a document "perfectly" answers a query, it has to mention all aspects expressed in the query. In other words, the query must be totally included in the document. Expressed through implication, this corresponds to the truth of $D \rightarrow Q$, i.e. $P(D \rightarrow Q)=1$. In most cases, $D \rightarrow Q$ is evaluated to an "uncertain truth" when one is not sure that every aspect of the query is mentioned in the document; or evaluated to an "imperfect truth" when only part of the query is mentioned in the document. These ideas are illustrated in the following examples, where we suppose that documents and queries are simple sets of unweighted index terms.

ideal case: D={information retrieval system, expert system}
Q={expert system}

non-ideal cases: D'=D
Q'={information system}

D"={database query, knowledge representation}
Q"={database query, data storage}

- In the ideal case, one is sure that the query Q is totally included in the document D. So $D \rightarrow Q$ is true.

- In the second case, one is not sure that the term "information retrieval system" includes the aspects of "information system", for certain people argue that "information retrieval system" is one particular case of "information system", and others consider them as being independent. The truth of D'→Q' is thus "uncertain".

- In the third case, document D" mentions a part of query Q", but not the whole query. The truth of D"→Q" is then "imperfect".

All these examples show, as is well-known, that if a document is to answer well a query, it must include every concept of the query.

Let us now look at another example. Suppose that two documents D and D' contain all terms of a query Q. D talks about the elements of the Q in detail and D' does not. If the documents are represented as a set of weighted index terms, we may have for example:

Q={t}
D={(t,0.9),...}
D'={(t,0.1),...}

If only a measure based on the inclusion of the query in the document is considered, i.e. D→Q, the same value is got for D→Q and D'→Q because D and D' do mention all aspects of Q. The difference between the two documents lies in the importance of term t in D and D', which reflects the importance of the document fragment related to the concept t in the original documents. If we name the earlier factor of judgement ***exhaustivity*** of the document to the query, we can name the current one ***specificity*** of the document to the query. In other words, the earlier factor says if all elements of the query are mentioned in the document; while the second says in what detail the elements of query are mentioned in the document. Expressed in notion of implication, D→Q corresponds to the exhaustivity, and Q→D corresponds to the specificity. If a query is wholly mentioned in a document, the D→Q is evaluated into 1; if the document concerns only the query, then the Q→D is 1.

In conclusion, if a function P is defined for measuring implications, the ***correspondance*** between query and document can be evaluated as follows:

$$R(D,Q) = F[\, P(D \rightarrow Q), P(Q \rightarrow D) \,] \qquad (2)$$

where F is a compromising function between the two implications.

Some remarks can thus be made about the terminology: "correspondence" is different from "pertinence" or "relevance" as defined in [SAL83]. The two latters are based on user's judgement, whereas "correspondence" is evaluated only by the system.

In many other models, these two elements are all included but somewhat mixed. In separating them, one is able to consider the system's characteristics in a more refined way. For example, the system can choose a compromise function between two factors according to the user's requirements.

The definitions of functions P and F strongly depend on system characteristics and user's requirements. Each model has its proper definitions, as shown in the next section. So no "universal" definition can be given here. Some discussions on this point will be carried out later.

IV. EVALUATION OF IMPLICATION

For the evaluation of D→Q, Van Rijsbergen [RIJ86] has given a logical uncertainty principle as is cited below:

> "Given any two sentences x and y; a measurement of the uncertainty of y→x relative to a given data set, is determined by the minimal extent to which we have to add information to the data set, to establish the truth of y→x."

We propose to consider and refine this idea of uncertainty using a particular logics: the modal logics ([HUG68] [ZEM75]).

The propositional logics permits us to express propositions like: "the propriety p is true for object a". Besides this, the modal logics adds two modalities: possibility and necessity. A necessary truth is one which can only be true; a possible truth is one which can be not true. One often distinguishes them by introducing the notion of "possible world": a necessary truth is true in all possible worlds; whereas a possible truth is true in a particular world but not in every possible world. The notion of "possible world" may be explained as: on a world (which may be a set of assertions), we make some changes for it to become another world, the second world is a "possible world" of the first. What is interesting in this formalism lies in the notion of "possible world", which allows us to formalize documents and queries in Information Retrieval.

If we construct an equivalent modal logic for the case of D→Q, the evaluation of the implication becomes the following:

Suppose that a query Q is a set of propositions, a document D is a set of assertions which compose an initial world. When Q is not satisfied in D (i.e. D→Q is not true), we (semantically) change some assertions in D (D will also be noted D_0) to transform it into D_1. We evaluate Q again in D_1. If D_1→Q is still not true, we change it into D_2, and so on, until D_n with D_n→Q is true. In this way, we have considered a succession of worlds $<D_0, D_2, D_3, ..., D_n>$ where D_i is a "possible world" of its precedent D_{i-1}. This succession of worlds can be equivalently expressed by a succession of changes of assertions $<C_1, C_2, C_3, ..., C_n>$, C_i being the change from D_{i-1} to D_i. In general, from a given D_i, we may have several possible semantic changes to the next world. So we can consider a tree of all document transformations illustrated in Fig.1.

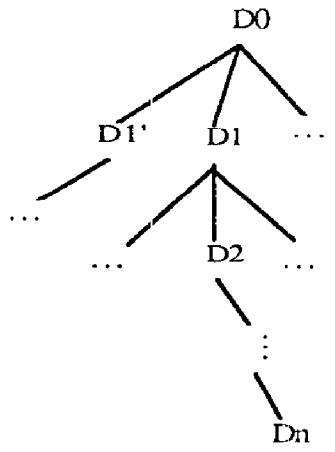


Fig.1

The world succession (or the change succession) correponds to one path from the root to a leaf. Among the possible paths from root to leaves, the one which has the minimal cost (i.e. the minimal change from the initial document) is chosen for the uncertainty measure of D→Q.

In this way, the uncertainty of D→Q can be measured by the distance from D to D_n: $d(D \rightarrow D_n)$. To obtain D_n, a sequence of changes is made on the initial document. So the basic operation is to measure the elementary distance from D_{i-1} to D_i, i.e. $d(D_{i-1},D_i)$. Once all elementary distances are evaluated, what we call the "mutual distance" from D_0 to D_n can then be evaluated in using an additive operator, for example, the addition of reals. If we take the addition, the mutual distance may be defined as:

$$d(D_0,D_n) = \sum_i d(D_{i-1},D_i)$$

More generally, the mutual distance may be expressed as:

$$d(D_0,D_n) = \bigoplus_i d(D_{i-1},D_i)$$

We would like now to consider a small change to Rijsbergen's uncertainty principle:

> Given any two information sets x and y; a measurement of the uncertainty of y→x relative to a given knowledge set K, is determined by the minimal extent E to which we have to add information to y, to establish the truth of (y+E)→x.

In this principle, K and E are not independent. E is relative to K in following way:

- E cannot be a subset of K, because K being defined in the system, this knowledge is available in any possible world. Thus if Q is not satisfied in a given possible world D_i, the additional knowledge needed must be found outside K.

- E can be inferred from K; in this case, the consideration of E does not increase the uncertainty (or does not decrease the certainty) of the truth of y→x;

- E can be something independent of K; this is the case of a pure addition of information which increases the uncertainty of the truth of y→x.

This principle of uncertainty corresponds very well to the ideas of global assumption and local assumption in modal logics ([HUG68] [ZEM75]), according to which an implication is expressed as:

$$K \mid y \rightarrow x \qquad (3)$$

where K is a "global" assumption, y is considered as a local assumption, and x a computed proposition. The global assumption is related to the model, so is true in any world. The local assumption is true only in the corresponding world. In real systems, the global assumption corresponds to the system's knowledge while the local assumption corresponds to documents.

The expression 3 covers more cases than expression y→x, for a system's factor is added so that it can be applied in any system. Correspondingly, a more general form for

formulae 2 is obtained:

$$K \mid R(D,Q) = F[\, K \mid P(D \rightarrow Q),\ K \mid P(Q \rightarrow D)\,] \qquad (4)$$

The above explanation concerns the evaluation of D→Q. The evaluation of Q→D is similar to it, for a query has been considered as a particular document description. Some slight modification will be necessary when one represents queries and documents differently. This will be discussed in VII.

V. ANOTHER WAY FOR IMPLICATION EVALUATION

In last section, a model based on modal logics is defined for the evaluation of D→Q. According to it the document (environment) is modified to find a representation which can match the query (proposition).

In fact, the world derivation during the evaluation of an implication may be considered in two ways:

- extending the environment to match the proposition; or
- reducing the proposition to match the environment.

Conceptually these two approaches based on modal logics are equivalent. The second one will be detailed in this section.

In contrast to the first approach, the proposition set is considered as a world. For the evaluation of D→Q, Q is progressively changed into Q_1, Q_2, ... until Q_n which is totally satisfied in D. The uncertainty of the initial implication D→Q can be measured by the distance from Q to Q_n. As in the first approach, several changes of Qi are possible at each instance. So a tree of possible derivation of Q is established. The minimal path from Q to Q_n is considered again to evaluate the uncertainty of implication. A corresponding principle for implication evaluation can be given as:

> For two information sets A and B, a measurement of uncertainty of A→B relative to a set of knowledge is determined by the minimal diminution from B to B' to establish the truth of A→B'.

VI. AN EXAMPLE OF CORRESPONDENCE MEASUREMENT

This example is developed to obtain part of a existing correspondence evaluation ([BOO80]). It is supposed that documents and queries are represented in set of independent weighted index terms. The distance from one world to another is defined as the proportion of reduced information in comparison with the initial quantity of information.

$$d(D_{i-1},D_i) = \frac{n(D_i - D_{i-1})}{n(D_0)}$$

$$d(Q_{i-1},Q_i) = \frac{n(Q_i - Q_{i-1})}{n(Q_0)}$$

The uncertainty of the implication is defined as the distance. The certainty is defined as the complement of the uncertainty. The function F in formulae 2 is defined as the product of two implications, i.e.:

$$R(D,Q) = F[P(D\rightarrow Q),P(Q\rightarrow D)] = P(D\rightarrow Q) * P(Q\rightarrow D)$$

Suppose that a document $D=\{(t1,a1),(t2,a2),(t3,a3)\}$, and a query $Q=\{(t1,b1),(t4,b4)\}$, where t_i is a term, a_i and b_i are respectively the weight of the term t_i in the document and the query (for simplicity we suppose $\Sigma a_i=1$, $\Sigma b_i=1$).

The Q is not totally satisfied in D, for the second part t4 is not explained in D. So the Q has to be reduced into $Qn=\{(t1,b1)\}$ in removing t4 from the query. In Qn there is only one the term t1 which is satisfied in D. The distance from Q to Qn is then:

$$d(Q,Qn) = b2/(b1+b2) = b2 = 1-b1$$

so $$P(D\rightarrow Q) = 1-d(Q,Qn) = b1$$

Similar to the evaluation of $D\rightarrow Q$, we get:

$$d(D,Dn) = (a2+a3) / (a1+a2+a3) = a1+a2 = 1-a1$$
$$P(Q\rightarrow D) = 1-d(D,Dn) = a1$$

Then, $$R(D,Q) = P(D\rightarrow Q) * P(Q\rightarrow D) = a1\cdot b1$$

When the query is a conjunction of two sub-queries $Q = Q1 \vee Q2$, it is considered that, if one of the sub-queries matches (in specificity and exhaustivity) a document, then the whole query matches the document. In the other words, we have the following inequations:

$$R(D,Q) \geq R(D,Q1)$$
$$R(D,Q) \geq R(D,Q2)$$

Then the correspondance of the query can be evaluated by:

$$R(D,Q) = MAX[R(D,Q1), R(D,Q2)]$$

Suppose a query as below:

$$Q = Q1 \vee Q2 = \{(t1,b1),(t4,b4)\} \vee \{(t2,c2),(t5,c5)\}$$
(where b1+b4=1 and c2+c5=1)

As in the above example, R(D,Q1) and R(D,Q2) are respectively evaluated into:

$$R(D,Q1) = a1\cdot b1$$
$$R(D,Q2) = a2\cdot c2$$

Then $$R(D,Q) = MAX(a1\cdot b1, a2\cdot c2)$$

These correspondence measurements are part of those in a fuzzy request model (see [BOO80]). All the evaluation equations in this model may be generated in a similar way from our model. This will not be show here.

VII. SOME DISCUSSIONS ABOUT EVALUATION OF IMPLICATIONS

VII.1. Evaluation of implications

In the equivalent modal logics, there is a succession of worlds which corresponds to semantic changes. The influence of the semantic changes on the initial world depends on the system's "intelligence" (i.e. K in formulae 3). One can compare a system with a human being. If he has some background in Artificial Intelligence, he knows that an "expert system" is a particular system of "artificial intelligence". If he has not, he may consider "expert system" as a concept completely independent of "artificial intelligence". It is the same thing for a system. The more a system has aquired knowledge, the more it has inference capabilities, the more it will be able to give precise answers. Incorporating this aspect, our system may then be viewed as in figure 2. This figure is a direct consequence of formula 3 and 4. If we consider a semantically independent model for Information Retrieval, the knowledge set is empty; whereas if the Information Retrieval model is semantically dependent, the semantic relations between elements are part of the knowledge set. This knowledge is often not that of the user, and because of this, the term "correspondence" is employed in the model rather than the well-known terms "relevance" or "pertinence". Only in the particular case where the system's knowledge corresponds to the user's knowledge, we may assimilate "correspondence" with "pertinence" or "relevance" as in [SAL83].

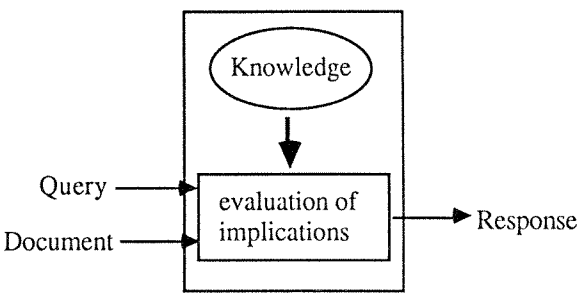


Fig.2

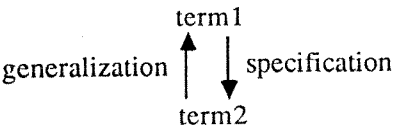
VII.2. The nature of documents and queries

In our former discussion, the same representation has been considered for documents and queries. In some real systems, queries and documents are represented in different data models. For example: documents are weighted while queries are not, or *vice versa.*

In this case, the evaluation of implication $D \rightarrow Q$ must be different to that of $Q \rightarrow D$, and formulae 1 becomes:

$$R(D,Q) = F[\ P(D \rightarrow Q),\ P'(Q \rightarrow D)\ ]$$

On the other hand, these implications are naturally different. For example, a query about "operating system" may be satisfied by a document about "Unix system"; but a query about "Unix system" can often not be satisfied by a document about "operating system". So during the evaluation of the two implications, the knowledge used is different. We would like to consider that the knowledge (semantic relation) is used in opposed ways. This is shown by the following example:



Term1 is more general than term2; term2 is more specific than term1. A query on term1 may be satisfied by a document on term2; a document on term2 may satisfy a query on term1. The semantic relation used in these assertions is the same, but in different ways.

VII.3. The function F

In many other models, the expression of R(D,Q) is given directly, without separating the two factors. In our formulae of R(D,Q), one is able to choose a function which favours one factor or another. In the vectorial model, the function F gives the same importance to the two factors. In the classical boolean model, F only concerns the implication D→Q (see next section).

In our opinion the choice of F could depend on the user's typology. If the user is an expert in the domain, his queries will more likely correspond to his real requirements. If the user is a beginner, he will not express his needs as well as an expert. Thus considering queries of an expert, we shall more probably retrieve documents containing every element of the query; while for a beginner, more probably, documents concerning only part of the query will be retrieved. So we are tempted to say that D→Q is more important for an expert than for a beginner and that Q→D is more important for a beginner than for an expert. (One can find an experiment on qualification of users in [DEF86].)

VIII. COMPARISON OF THE MODEL WITH SOME EXISTING MODELS

It is partially shown in VI that our model can be related to the fuzzy model. In this section, some more comparisons will be made with other Information Retrieval models.

VIII.1. The vector retrieval model

In this model, a document is represented by a n-dimensional vector of index terms (or keywords) as follows:

$$D = (a_1, a_2, ..., a_n)$$

where n is the number of attributes defined in the system, and a_i represents the weight (between 0 and 1) of term A_i attached to D. Then a document set can be represented by a matrix:

$$\begin{array}{c|cccc|} & A_1 & A_2 & \dots & A_n \\ D_1 & a_{11} & a_{12} & \dots & a_{1n} \\ D_2 & a_{21} & a_{22} & \dots & a_{2n} \\ \dots & & \dots & & \\ D_m & a_{m1} & a_{m2} & \dots & a_{mn} \end{array}$$

A query Q is also represented by a set of possibly weighted terms: $(b_1,b_2,...,b_n)$.

During retrieval processing, the system selects the documents D_i which give the highest similarity with the query - $Sim(Q, D_i)$. Some well-known definitions of the similarity function are given below:

$$Sim(Q, D_i) = \frac{2\sum_j (a_{ij} \cdot b_j)}{\sum_j a_{ij} + \sum_j b_j} \qquad (5)$$

$$Sim(Q, D_i) = \frac{\sum_j (a_{ij} \cdot b_j)}{\sum_j a_{ij} + \sum_j b_j - \sum_j (a_{ij}.b_j)} \qquad (6)$$

$$Sim(Q, D_i) = \frac{2\sum_j a_{ij} \cdot b_j}{[\sum_j (a_{ij})^2 \cdot \sum_j (b_j)^2]^{1/2}} \qquad (7)$$

These definitions can be easily expressed in terms of exhaustivity and specificity. Consider the transformation of (6) as an example:

For P and F in formulae 2, we give the following definitions:

$$P(D \rightarrow Q) = \frac{n(D \cap Q)}{n(D)} \quad \text{and} \quad P(D \rightarrow Q) = \frac{n(D \cap Q)}{n(Q)}$$

where $n(D_i)$ is $\sum_j a_{ij}$, $n(Q)$ is $\sum_j b_j$, $n(D_i \cap Q)$ is $\sum_j a_{ij} \cdot b_j$.

So

$$R(D_i, Q) = \frac{\sum_j a_{ij}.b_j}{\sum_j a_{ij} + \sum_j b_j - \sum_j a_{ij}.b_j} = \frac{1}{1/P(D_i \rightarrow Q) + 1/P(Q \rightarrow D_i) - 1}$$

$$= \frac{P(D_i \rightarrow Q) \cdot P(Q \rightarrow D_i)}{P(D_i \rightarrow Q) + P(Q \rightarrow D_i) - P(D_i \rightarrow Q) \cdot P(Q \rightarrow D_i)} = F[P(D_i \rightarrow Q), P(Q \rightarrow D_i)]$$

VIII.2. The boolean model

Only the classical boolean model (not the extended boolean model) is considered here. It is also assumed that documents are represented by a set of index terms; queries being logical combinations of these terms. A document is to be retrieved if the terms in the document satisfy the boolean expression of the query.

This corresponds to the notion of exhaustivity in our model: D is a good response to Q if D→Q is true, i.e. P(D→Q)=1. Another factor of the formulae1 is not taken into account. (This assumes that, when a index term appears in a document, it is all of mportance; when a term is absent, the document does not concern it at all.) Thus the boolean model is a restricted case of our general model:

$$R(D,Q) = P(D \rightarrow Q)$$

VIII.3. The probabilistic model

Here is considered the standard probabilistic model. Given a query Q, the probability of relevance of a document D is P(rel|D). In Bayes' thoerem ([RIJ79]), it is evaluated by the following formulae:

$$P(\text{rel} \mid D) = \frac{P(D|\text{rel})\, P(\text{rel})}{P(D)}$$

where P(D)=P(D|rel)P(rel) + P(D|nrel)P(nrel)

After each retrieval operation, the user revises the value of P(D|rel). This revision is based on the user's judgement of the document relevance. In fact, the judgement of the user can be composed of two elements - those that we name "exhaustivity" and "specificity". So the two implications are implicitly included in the evaluation of probability of relevance.

IX. CONCLUSION

In this paper, a general model is proposed, which consists in a development of Rijsbergen's approach using modal logics. Although this model cannot be described here in a great detail, we would like to argue that it can be applied to define all existing models; for in any information retrieval system, there are the two aspects - exhaustivity and specificity. This model, however, remains to be further investigated before any practical application.

ACKNOWLEDGEMENT

The author is glad to take this opportunity to thank Pr. Yves Chiaramella for his careful reading of this article and his many helpful suggestions.

REFERENCES

[BOO80] Bookstein A.
Fuzzy Requests: An approach to Weighted Boolean Searches, *Journal of the American Society for Information Science* , July 1980

[CHE82] Chemdata Sarl
Manuel d'Utilisation Texto, 1982, Lyon

[DAB75] Dabrowski M.
A General Model of Distribution of Objects in Information Retrieval Systems, *Information Systems*, Vol.1, Pergamon Press, 1975

[DEF86] Defude B
Etude et Réalisation d'un Système Intelligent de Recherche d'Informations: Le Prototype IOTA, Thèse INPG, 1986

[HUG68] Hughes G., Gresswill M.
An Introduction to Modal Logic, Methuen, 1968

[NLM79] National Library of Medicine
MEDLARS, The Computerized Literature Retrieval Service of the National Library of Medicine, Department of Health, Education and Welfare, Publication NIH 79-1286, Jan 1979

[RIJ79] Van Rijsbergen C.J.
Information Retrieval, 2nd edition, London, Butterworths, 1979

[RIJ86] Van Rijsbergen C.J.
A Non-classical Logic for Information Retrieval, *Computer Journal*, Vol.29(6), 1986

[SAL71] Salton G. (editor)
The SMART Retrieval System - Experiments in Automatic Document Processing, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971

[SAL83] Salton G., McGill M.J.
Introduction to Modern Information Retrieval, International Student Edition, 1983

[WAL79] Waller W.G., Kraft D.H.
A Mathematical Model of a Weighted Boolean Retrieval System, *Information Processing & Management*, Vol.15, Pergamon Press Ltd., 1979

[ZEM75] Zeman J.J.
Modal logic, Oxford, 1975