Probability, information and IR

Gianni Amati*
Fondazione Ugo Bordoni
v.B. Castiglione 59
00142 Rome, Italy
email: gba@fub.it

Keith C.J. van Rijsbergen
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, Scotland
email: keith@dcs.gla.ac.uk

1 Introduction

We investigate probability theory in modeling retrieval of information. The definition of retrieval mechanisms based on the probability theory needs first to make clear what is meant by a probability space in the context of IR.

When we talk about probabilities, we essentially deal with measuring a boolean or a sigma algebra of events. A probability space is thus a triple $< \Omega, \mathcal{A}, \mathcal{P} >$ where Ω is the *basic space*, $\mathcal{A}$ is a sigma algebra on Ω , which is an algebra closed with respect to complementary set operation, intersection and union of countable sets, and P a measure function on the algebra $\mathcal{A}$ such that $P(\Omega) = 1$. The elements of Ω, in the language of probability theory, are also called *elementary events* or *outcomes* while in set theory they are called *atomic* elements. We may wonder what are the possible explanations for the entities Ω, $\mathcal{A}$ and P in the context of IR. In general, they depend on the nature of the concept of relevance, whether it is statistical, semantic, or subjective. As far as we see there are four ways to look at the relevance. The first is purely statistical: the ranking of documents is performed by exploiting frequencies of words within a document and with respect to the entire collection of documents. The information need of the user is expressed by a possibly weighted vector of keywords. Salton's vector space model

*Work carried out in the framework of the agreement between the Italian PT Administration and the Fondazione Ugo Bordoni

belongs to this class . The second approach considers relevance as a purely subjective feature of retrieval: van Rijsbergen's probabilistic model falls into this class. Our approach, which we introduce in this article, shows how to define a foundational theory which combines both statistical and subjective approaches. The semantic approach has his roots in the semantic theory of Carnap [6]. Probabilities are assigned to the sentences of the language and the information content of a sentence is the amount of its uncertainty. However Carnap's theory measures the information content by essentially counting the number of possibilities that sentence excludes. Hence, the content measure of a sentence α, $content(\alpha)$, increase as soon as the expressiveness of the language increases. However, the qualification "semantic" for the information has been used in a wider extension. There is an intensional connection between sentences, which can be thought in terms of conditional logic or in probabilistic relation of causation [8, 7]. In this paper we explore the more restricted meaning of the "semantic" point of view. More formally we may enumerate the following four main probabilistic frameworks:

1. Terms are elementary events: the probability space is a quadruple $< T, P(T), Q, M >$ where $P(T)$ is the power set of the basic space T , Q is the query considered as a measure function and M is the document–by–term matrix. This matrix allows us to consider documents as stochastic variables and to define thus a probability measure E_Q derived by the conditioning expectation on the measure Q.

2. The set D of documents is the set of elementary events. The measure space is the quadruple $< D, P(D), R, M^* >$ where $P(D)$ is the power set of D. R is a measure of document relevance defined by the user, M^* is the term–by–document matrix. This approach uses the dual perspective of the previous space: documents are stochastic variables and define a probability measure E_R as the conditioning expectation on the measure R.

3. The algebra is isomorphic to the Lindembaun algebra $\mathcal{L}_{\mathcal{R}}$ generated by the boolean logical operators from the set of terms T considered as propositional constants. The set D of documents are boolean conjunctions of terms hence are treated as elementary events. An event Rel (the "relevance " event), is also added to the algebra $\mathcal{L}_{\mathcal{R}}$. In contrast to the previous space the matrix M^* as well as the R are considered binary.

4. Terms and documents are possible worlds in the semantics of modal logic and the probability may be computed by different techniques: one is imaging on conditional logic.

We show that the first approach leads to the Salton's vector space model. The probability space on the algebra $\mathcal{L}_{\mathcal{R}}$ is instead taken by the probabilistic model of van Rijsbergen. We will show that this model is subsumed by the $< D, P(D), R, M^* >$ model by considering a particular relevance measure distribution. In fact if R is the counting measure of relevance and if the document-term matrix is binary (i.e. takes only the occurrence and non-occurrence values 0 and 1) then we obtain the same term-weighting formula of the probabilistic model.
The probability space $< D, P(D), R, M^* >$ is indeed more general, as:

1. the values of R can be arbitrarily defined by the user;

2. these relevance values can be easily integrated with any document-term matrix. The model $< D, P(D), R, M^* >$ seems therefore to be a way for integrating the user feedback with the frequency analysis on the collection D of documents deeper than that based on occurrence/ non-occurrence of terms and relevance/non-relevance of the probabilistic model.

1.1 A simple probabilistic retrieval model: the Salton's vector space model

The first approach in IR leads to a standard application of measure/ probability theory. The algebra A is that generated by using the set T of terms of the language as the basic space. The events of the algebra A is the power set $P(T)$ of T, that is the set of all subsets of T: an atomic element is any singleton $\{t\}$ corresponding to the element $t \in T$.
Given a query q, which is a set of terms, we can associate a probability (or a measure) function Q (in general the set of terms not occurring in q is a set of null measure, as in Example 1 below). We assume that a document x is a random variable from the set T of terms in the set of real numbers (x is a measurable mapping if Q is a measure). In the following x_i denotes the value of x on the i-th term t_i and q_i denotes $Q(t_i)$.
Since the measure space is discrete the Stieltjes integral $Q(T) = \int_T dQ$ turns out to be $\sum_{t \in T} Q(t) = \sum_{t \in q} Q(t)$ ($= 1$ if Q is a probability measure). The

expectation $E(x)$ on the random variable x is: $\int_T x dQ = \sum_{t_i \in q} \int_{t_i} x dQ$. Now $\int_{t_i} x dQ$ is $x_i q_i$ hence::

$$E_Q(x) = E(x|Q) = \frac{\int_T x dQ}{\int_T dQ} = \frac{\sum_{t_i \in T} x_i q_i}{Q(T)} \tag{1}$$

Proposition 1.1 *Let $< T, P(T), Q >$ be a measure space. The expectation of the random variable x turns out to be the similarity function of the vector space model namely the inner product of x and Q considered as vectors.*

1.2 A retrieval model by using continuous or fuzzy values for user feedback

Differently from the previous section we assume that a term t is a random variable on the set of documents. $a^{\ i}_{doc}$ denotes the value $doc(t_i)$ of a document doc on the term t_i. This last value is in general computed on the basis of word frequency analysis. The value $t_i(doc)$ of a term t_i on a document doc is obtained by the inverted file of the document-term matrix $a^{\ i}_{doc}$:

$$t_i(doc) =_{def} doc(t_i) =_{def} a^{\ i}_{doc}$$

Definition 1 Let R be any relevance measure on the set of documents. For convenience we assume that R takes the value $r_{doc} = R(doc)$ with $0 \leq r_{doc} \leq 1$ on each document doc according to the user's relevance assessment.

Notice that $R(D) = R(\bigcup_{doc \in D} doc) = \sum_{doc \in D} R(doc) = \sum_{doc \in D} r_{doc}$ The expectation $E(t_i|R(D))$ on the random variable t_i is the Stieltjes integral on the measure (or probability) R. $\int_D t_i dR = \sum_{doc \in D} \int_{doc} t_i dR$ and $\int_{doc} t_i dR$ is $a^{\ i}_{doc} r_{doc}$, hence:

$$E(t_i|R(D)) = \frac{\int_D t_i dR}{\int_D dR} = \frac{\sum_{doc \in D} a^{\ i}_{doc} r_{doc}}{\sum_{doc \in D} r_{doc}} \tag{2}$$

$E(t_i|R(D))$ is the *expected probability of relevance of the term* t_i and will be denoted by $E_R(t_i)$ or $E(t_i|R)$.

Analogously to the case of computing the relevance measure of a term, we may introduce the *non-relevance* measure $\neg R$, defined as $\neg R(doc) = 1 - R(doc)$. If a subset X of D has m elements then $\neg R(X) = m - \sum_{doc \in X} r_{doc} =$

$m - R(X)$. We then define a probability measure as the *expected measure of non relevance of the term* $t_i E_{\neg R}$

$$E_{\neg R}(t_i) = E(t_i|\neg R(D)) = \frac{\int_D td\neg R}{\int_D d\neg R} = \frac{\sum_{doc \in D} a^{\ i}_{doc} \neg R(doc)}{\neg R(D)} \quad (3)$$

1.3 Applying decision theory:refining the expected relevance measure of a term

When we decide that a term or document is relevant or not on the basis of the expected measures of relevance and non-relevance, an error can occur and then a loss is produced. The *expected monetary value (EMV)* on taking a relevance/non relevance decision can be represented by the formula:

$$EMV = \lambda_1 * E_R(t) + \lambda_2 * E_{\neg R}(t) \quad (4)$$

or by the formula:

$$EMV1 = log \frac{\lambda_1 * E_R(t)}{\lambda_2 * E_{\neg R}(t)} \quad (5)$$

Let us assume that the user has the chance to choose just one term at a time for formulating a query and that she is trying to decide whether to use the term t or not. t has a priori a relevance probability $E_R(t)$. Suppose also that t is relevant to the information need of the user. λ_1 would be then the 'gain' if she decides t while λ_2 would be the 'cost' if t is not the case (this has the a priori probability $E_{\neg R}(t)$). Thus the decision on the relevance /non-relevance of t should take into account that

$$\begin{aligned} EMV &> 0 \\ EMV1 &> 0 \end{aligned} \quad (6)$$

If 't is relevant', then she will gain the amount of information of non-relevance of t, that is $Inf_{\neg R}(t|D)$. On the other hand, the loss λ_2 can be quantified by the amount of information of relevance of t, that is $Inf_R(t|D)$. These values are given by the Carnap-like clause $Inf(x) =_{def} 1 - E_P(x)$, that is

$$Inf_{\neg R}(t|D) = 1 - E_{\neg R}(t|D) = \neg R(1|D) - \int_D (t|D) d\neg R = E_{\neg R}(\neg t|D)$$
$$Inf_R(t|D) = 1 - E_R(t|D) = E_R(\neg t|D)$$

Substituting the λ 's into (7), and using the *log* we have:

$$w(t_i) = log \frac{E_R(t_i)}{E_{\neg R}(t_i)} > 0 \tag{7}$$

$$w(t_i) = log \frac{E_R(t_i) * E_{\neg R}(\neg t_i)}{E_R(\neg t_i) * E_{\neg R}(t_i)} > 0 \tag{8}$$

1.4 An example: the probabilistic model of van Rijsbergen

Let us apply the model $< D, P(D), R >$ with a particular relevance measure R. We assume

1. R is the counting measure for the relevance of documents i.e. R takes a value $r_{doc} = 0$ or 1 on each document according to the user relevance feedback;
2. $a^{\ i}_{doc}$ is the *counting document-term matrix*, that is

$$a^{\ i}_{doc} = \begin{cases} 1, & \text{if the i-th term occurrs in doc;} \\ 0, & \text{otherwise.} \end{cases}$$

In the following n_R denotes the cardinality of the relevant set of documents, N that of D, r^i that of the set of relevant documents in which the term t_i occurs, $n^{\ i}_{\neg R}$ that of the set of non relevant documents in which the term t_i occurs, n^i that of the set of documents in which the term t_i occurs.

By definition of $a^{\ i}_{doc}$, the value $\sum_{doc \in D} a^{\ i}_{doc} r_{doc}$ is the cardinality r^i of the set of relevant document in which the term t_i occurs. Substituting r^i into (2) we get the expectation $E(t_i|R(D))$ on the random variable t_i as $E_R(t_i) = E(t_i|R(D)) = \frac{r^i}{n_R}$

Analogously, $E_{\neg R}(t_i) = \frac{\sum_{doc \in D} a^{\ i}_{doc} \neg R(doc)}{\neg R(D)} = \frac{n^i - r^i}{N - n_R}$. Finally we obtain $E_R(\neg t_i) = 1 - E_R(t_i) = \frac{n_R - r^i}{n_R}$ and $E_{\neg R}(\neg t_i) = 1 - E_{\neg R}(t_i) = \frac{N - n_R - n^i + r^i}{N - n_R}$

The weight $w(t_i)$ is the well known weighting formula of the probabilistic model:

$$w(t_i) = log \frac{E_R(t_i) * E_{\neg R}(\neg t_i)}{E_R(\neg t_i) * E_{\neg R}(t_i)} = log \frac{\frac{r^i}{n_R - r^i}}{\frac{n^i - r^i}{N - n_R - n^i + r^i}} > 0 \tag{9}$$

Proposition 1.2 *Let R be the counting relevance measure on the set D of documents and (a^i_{doc}) the counting document-term matrix. Consider the measure space $< D, P(D), R >$. The measure of absolute relevance $w(t)$ of a term t, defined in* (8)*, is the term weight of the van Rijsbergen's probabilistic model.*

2 Acknowledgment

Work carried out in the framework of the agreement between the Italian PT Administration and the Fondazione Ugo Bordoni.

References

[1] Keith C.J. van Rijsbergen, *Information Retrieval*, Butterworths, Sydney, 1979.

[2] Afred Renyi,*Foundations of Probability*, Holden-Day Press, San Francisco, 1969.

[3] W. Bruce Croft, Relevance Feedback and Inference Networks,*Communication of the ACM*, SIGIR-93, 2-11,1993.

[4] James R.Chen and Nathalie Mathe, A contextual Relevance Model for Adaptive Information Retrieval,in *Indexing and Reuse in Multimedia Systems*, Proceedings of the AAAI-94 Workshop, 175-186,1994.

[5] W.S. Cooper, F.C. Gey, and D.P. Dabney, Probabilistic retrieval based on staged logistic regression, *SIGIR-92*, 198-210, 1992.

[6] Y. Bar-Hillel, R. Carnap, Semantic Information, *British J. of the Philosophy of Science*, 4, 147–157, 1953.

[7] P. Gärdenfors, *Knowledge in Flux*, MIT Press, Cambridge MA, 1988.

[8] J. Hintikka, On semantic information, in J. Hintikka and P. Suppes (eds), *Information and Inference*,Synthese Library, Reidel, Dordrecht, 1970.