From a Qualitative towards a Quantitative Representation of Uncertainty in a Situation Theory based model of an Information Retrieval System

Mounia Lalmas
Department of Computing Science
University of Glasgow
Scotland

Introduction

We have advanced a model [4] of an information retrieval (IR) system based on van Rijsbergen's postulate [8] that relevance could be modelled by an implication. Given that d and q are the sentences in a logic that formalise, respectively, the *information content* of the document and the *information need* phrased within the query, the evaluation of $d \rightarrow q$ would express the relevance[1] of the document to the query. Several approaches have been proposed to evaluate this implication $d \rightarrow q$; ours' is based on the *Transformation Principle* [4]:

> Given a document d, a query representation q; the evaluation of $d \rightarrow q$, relative to some knowledge set K, is determined by the minimal transformation applied to d, to obtain some d' such that d' contains q.

First, an initial representation of the document's information content (e.g., *d*) is acquired, which typically comprehends the information *explicit* in the document. The information content can grow as the initial representation is transformed into successive, less partial, representations (e.g., *d'*), thus incorporating the information *implicit* in the document.

The notion of transformation is not new [2]; the novelty of our approach is to base the transformation on the *flow of information*. A document is relevant to a query if either the information need phrased in the query is explicit; or the document is also relevant, maybe to a lesser extent, if the information need is implicit in the document. The information implicit in the document is induced from the flow of information that stems from the information explicit in the document. Therefore, an appropriate model of an IR system should capture the flow of information.

To adequately express the flow of information, we use a theory of information and its flow that was developed by Barwise, namely Situation Theory [1, 3], to construct our model. We have formalised the transformation of a representation with the so-called situations, types and constraints, thus contributing to a truthful representation of information and its flow in an IR system.

Some documents are often more relevant to a query than other documents, so retrieved documents should be ranked according to their relevance to the query. Our ranking of the document was based upon the fact that the flow emanating from a document's initial representation can be uncertain. For example, in a document that refers to the Canadian Prime Minister, if we know that the document refers to the year 1994, then we also know implicitly that Canadian Prime Minister refers to J. Chretien. However, if the document refers to the years 1975 until now, we cannot affirmatively say that Canadian Prime Minister refers to J. Chretien. The flow of information in that case is uncertain.

---

1 Here we are not talking of the user-relevance judgement, but the system-relevance assessment.

The *uncertainty of the flow* can then be used to express the relevance of the document to the query. The more uncertain the flow, the less relevant the document.

Situation Theory offers a *qualitative* representation of the uncertainty. This paper shows that a *quantitative* representation of the uncertainty can elicit from the qualitative one, thus enabling the formulation of the ranking of documents according to their relevance to a query. Our goal is not to describe in details one approach, but simply to show that this démarche is feasible; simple examples will be used to illustrate our explications.

Information and its flow in IR

Situation Theory takes its foundation from the work of Dretske [4]. There, perception is defined as the process by which information is delivered in an analog form to a cognitive agent for its selective use. Cognition is described as the conversion of the information a cognitive agent receives in analog form into digital form. The conversion, referred to as digitalisation, involves a loss of information because it turns a structure of greater information content to one of lesser information content. Indeed, the perception process often embodies more information than can be extracted and/or exploited by the cognitive agent. Only some of the information the perception process carries in analog form is retained.

The indexing process in IR can be compared to a digitalisation process. The IR system is the cognitive agent. The document is an object that contains information in analog form. The information that is digitalized constitutes the document's initial representation, which is not an exhaustive description of the information content of that document. This representation customarily depicts the explicit information content of the document, which is determined at most from the meaning of the sentences of the document.

Information and meaning are two different concepts; often, the information exceeds its meaning. For example, the statement "Keith is at home" means that Keith is indeed at home. It does not mean that "Keith is at home and not at work" though Keith being at home implies that Keith is at home and not at work. The statement "Keith is at home" carries the information "Keith is at home and not at work". This information follows from the flow of information that springs from the meaning attached to the statement. Therefore, understanding the *meaning* attached to the sentences of a document is important, but is insufficient for determining the *information content* of the document. A more thorough representation of a document's information content needs the appropriate capturing of the flow of information. This is possible with the use of Situation Theory.

Situation Theory

Situation Theory [1, 3] is a theory of information which, by following a mathematical approach, models more naturally the way humans handle information and its flow. It can be compared to Quantum Mechanics, where an ideal representation of an electron is adopted, even if it is not well understood what an electron is. Hence, it becomes possible to model the behaviour and interaction of electrons. Situation Theory adopts the same approach. It proposes first to represent the notion of information without really indicating what information is; from this, a model of the flow of information is derived.

In Situation Theory, an item of information is considered as an entity and is represented by an *infon*. For example, the information concerning the fact that Mounia is swimming is modelled by the infon *«Swimming, Mounia; 1»*. If the information was that Mounia is not swimming, the corresponding infon would be *«Swimming, Mounia; 0»*. Nothing is said so far about the truth or falsity of the two infons above; they are representations of information items. What makes them true is any *situation* which pertains to them. If we call s the concerned situation, we write $s \models \ll Swimming, Mounia; 1 \gg$ to say that s *supports* the corresponding infon.

The two infons *«Swimming, Mounia, sea, 11am; 1»* and *«Swimming, Mounia, pool, 3pm; 1»* have in common the information that Mounia is swimming. What differs is the place and the time of the action. This similarity is represented in the theory by *types*. Here the corresponding type would be $\varphi = [\dot{s} \mid \dot{s} \models \ll Swimming, Mounia, \dot{p}, \dot{t}; 1 \gg]$, which classifies all the situations where a person referred to as Mounia is swimming at a given time and place[2]. Any situation that supports the information that Mounia is swimming at a given time and place is said to be *of type* φ; we write $s \models \varphi$ as well.

The information represented by the two types $\varphi = [\dot{s} \mid \dot{s} \models \ll present, smoke, \dot{p}, \dot{t}; 1 \gg]$ and $\psi = [\dot{s} \mid \dot{s} \models \ll present, fire, \dot{p}, \dot{t}; 1 \gg]$ is *informationally* related because the fact that there is smoke in a place carries the information that there was (or is) a fire nearby. There is a flow of information that indicates that a situation of type φ additionally carries the information that some other situations are of type ψ. The relationship between the two types is modelled by a *constraint*, written $\varphi \rightarrow \psi$. The application of this constraint to a situation s_1 is possible if first $s_1 \models \varphi$, and then informs the existence of a situation[3] s_2 such that $s_2 \models \psi$; in other words, $s_1 \models \varphi$ carries the information that $s_2 \models \psi$.

Constraints do not always hold. For example, the above constraint $\varphi \rightarrow \psi$ may not hold if the place where the smoke is seen is London (it was not smoke, but smog). The constraints that always hold are called *unconditional* and those that do not are called *conditional*. A conditional constraint is associated with *background conditions*, and is written $\varphi \rightarrow \psi \mid B$ to highlight the fact that the constraint $\varphi \rightarrow \psi$ holds if the background conditions captured within B are met by the situation to which the constraint is applied. If we interpret the background conditions as a set of types, the satisfaction of the background conditions B by the situation s_1 such that $s_1 \models \varphi$ involves three cases:

- $s_1 \models B$; the situation s_1 satisfies all the background conditions. The use of the constraint is certain as well as the resulting flow. The information supported by s_2 (i.e., $s_2 \models \psi$) is certain with respect to s_1.

- $s_1 \not\models B$; at least one of the background conditions is not satisfied. The constraint cannot be used.

- $s_1 \models ?B$; it is not known whether some of the background conditions are satisfied by s_1. The use of the constraint is uncertain, and the resulting flow is uncertain. The information supported by s_2 (i.e., $s_2 \models \psi$) is uncertain with respect to s_1.

Situations and types show similarities with documents and their information content. Constraints are the perfect tool to represent thesaural, or any semantic or pragmatic relationships which constitute the knowledge set upon which the transformation process is based. A document, if viewed as a situation, supports information that corresponds to the explicit information content (digitalized) of the document; the document often carries implicit information that depends on the available constraints, some of which are uncertain. The uncertainty that arises from these flows is used to assess the relevance of the document to the query.

---

2 Here $\dot{s}$, $\dot{p}$ and $\dot{t}$ are parameters, denoting a situation, a place and a time, respectively.

3 s_2 can be either the same as s_1 or different because a situation can carry information about itself as well as other situations.

The proposed model

The initial representation of a document's information content is modelled as a situation d, and a query, which expresses an information need, is modelled as a type φ. The knowledge set is modelled by constraints. If $d \models \varphi$ then the document is relevant to the query because φ is part of the explicit information content of the document. Otherwise, constraints are used to find a flow that could transform d to a situation that supports φ. Suppose that we have $d \models \psi$ and $\psi \to \varphi$ is a constraint of the knowledge set. We have three cases:

- If $\psi \to \varphi$ is conditional and the background conditions are not satisfied by d, then this constraint cannot be used to determine a flow that would lead to the information need.
- If the constraint $\psi \to \varphi$ is either unconditional, or conditional but such that d satisfies the background conditions, then the document initially represented by d is relevant to the query because d can be transformed to some d' such that $d' \models \varphi$.
- If $\psi \to \varphi$ is conditional and it is not known whether the background conditions are satisfied by d then the document *may be* relevant to the query because d could be transformed into a situation d', such that $d' \models \varphi$.

In the third case, the flow of information is uncertain. The relevance of the document is strongly dependent on this uncertainty because the more uncertain is the flow, the less relevant the document is to the query. This uncertainty is engendered by the use of uncertain conditional constraints. We show next that a quantitative representation of this uncertainty can be derived from the background conditions associated to these constraints. The quantitative expression of this uncertainty is used to rank documents in order of the relevance, so it should be defined such that the following properties are verified:

- obviously, the more uncertain is a flow, the less relevant the document.
- a *sequence* of transformations may be necessary to obtain a representation that contains the information requested by the query. The more transformations are necessary, the less relevant the document is.
- there may be alternative ways to transform a document. The alternative transformations that lead the initial representation to the same representation are referred to as *parallel* transformations. The more such transformations, the more relevant the document is.

A ranking of documents based on background conditions

Although it may not be know whether a situation satisfies some given background conditions, the situation could be closer to satisfy these background conditions than other situations. Consequently, the uncertainty involved in the transformation of the former situation should be less than with the transformation of the latter situations. A ranking of documents can be based on the extent to which the situation modelling the initial representation of a documents satisfies the background conditions of the constraints used in the flows that bears the information need.

The background conditions can be particularly complex to identify. In every day reasoning, people often use background conditions, though they are not aware of them. As Devlin [3] points out, background conditions become a concern only when a constraint fails. In the literature, no precise definition of the background conditions is advanced. We propose two definitions[4]:

- **(I1)** how many conditions are satisfied?

Let B be the set of background conditions sating that "it is snowing" and "the city is Montreal" (for simplicity, the information items are not represented by types). Let s be a situation that supports the information that it is snowing, and let s' be a situation where it is not known whether it is snowing or the city is Montreal. The situation s is closer to satisfy the background conditions in B than is the situation s'.

- **(I2)** how many alternatives are there?

Suppose that information items correspond to words which may be polysemic (they have several senses), and that constraints model synonymy. Let B be the background conditions attached to a constraint stating that the word w' is a synonym of the word w if the sense of w is 1. If the word w has two senses in a situation s and three in the situation s' then w is more ambiguous in the latter than in the former situation. If a constraint is defined for each sense of the word w (w has synonyms for each of its senses), then the situation s is closer to satisfy B than is the situation s' because the number of *alternative* flows to the one defined by the background conditions B in the situation s' is higher than from that word in the situation s.

The above two definitions of the satisfaction of background conditions can lead to the construction of a function N that can be used to rank situations according to the extent they satisfy background conditions. With definition I1, a formulation of N can be based on the number (ratio) of the background conditions satisfied by a situation. The higher this number for a given situation s and background conditions B, the higher is $N(s,B)$. In the first above example, since s satisfies more conditions in B that s' does, we have $N(s,B) \geq N(s',B)$. If the background conditions B are all satisfied by a situation s the value $N(s,B)$ is the highest, for no uncertainty is introduced.

With definition I2, the function N can be defined as inversely proportional to the number of alternative transformations there is with respect to some situation s and background conditions B. In the second above example, there are more alternative transformations from the situation s' than there are from the situation s, so we have $N(s,B) \geq N(s',B)$. If no alternative transformation occurs from a situation s with respect to some background conditions B, the value $N(s,B)$ is also the highest.

A measure of relevance can be based on the function N. Let s be the situation that models the initial representation of a document's information content. Suppose that a single transformation involving one constraint $\varphi \to \psi | B$ is necessary to reach a representation of the document's information content that contains the information need. The higher $N(s,B)$, the higher the relevance of the document. If no additional criteria is taken into account (e.g., the weight of information), the relevance of the document to a query can be itself $N(s,B)$.

Sequential transformations might be necessary to reach the information need, or parallel transformations might occur in reaching this information need. If a measure of relevance can be based on the function N, then the values of the function N should decrease with the number of sequential transformations, and increase with the number of parallel transformations.

4 There are many other definitions; for example, the satisfaction of an individual background condition can itself be uncertain. The above two definitions are simple, but sufficient for our purpose.

Suppose that a situation s is transformed into a situation s', which is then transformed into a situation s''. Let us assume that the background conditions involved in the two transformations are B and B', respectively:

$$s \xrightarrow{\quad B \quad} s' \xrightarrow{\quad B' \quad} s''$$

We define the sequence of background conditions as $B;B'$ to represent that the first transformation depends on B and that the second depends on B'[5]. In definition I1, the number of background conditions to be satisfied (at least) augments with each additional sequential transformation. Obviously, the more conditions there are to be satisfied, the more uncertain becomes the resulting flow, so $N(s,B) \geq N(s,B;B')$. In definition I2, sequential transformations imply that the number of alternative transformations with respect to s and B augments, so $N(s,B) \geq N(s,B;B')$. With both definitions I1 and I2, the function N captures that the more sequential transformations are required, the more uncertain is the information supported by s'' with respect to s.

Suppose a situation s can be transformed in two (parallel) ways into the same situation s', with respective background conditions B and B' (each transformation can itself be a sequence of transformations). Parallel transformations can be viewed as accumulation of evidence towards the information supported by s' with respect to s:

$$s \quad \begin{array}{c} B \\ \xrightarrow{\qquad\qquad} \\ \xrightarrow{\qquad\qquad} \\ B' \end{array} \quad s'$$

We define the parallel combination of background conditions as $B \| B'$. A situation s satisfies $B \| B'$ if s satisfies B or B' (the satisfaction may also be with respect to other intermediary situations if sequential transformations are involved). With definition I1, the number of satisfied background conditions in $B \| B'$ is the smallest of the numbers of satisfied background conditions in B and B', separately. The transformation of s into s' may become more certain as the number of parallel transformations increases, for one of them may be based on background conditions that are all satisfied. Consequently, $N(s,B) \leq N(s,B \| B')$. With definition I2, the number of alternative transformations with respect to $B \| B'$ decreases since there is a convergence towards the same situation s', so $N(s,B) \leq N(s,B \| B')$. Again, with both definitions, the function N captures that the more parallel transformations, the higher the value of N.

To recap: N is a quantitative representation of the uncertainty of the flow of information which is qualitatively captured by background conditions. With our use of the Transformation Principle, the uncertainty of the flow leading to a transformation is used to express the relevance of a document to a query. N can be used as a basis in the formulation of that relevance. A ranked output of retrieved document according to their relevance degree can be obtained. The function N can be simple or more sophisticated, and includes other information such as weight of information, dependence of the background conditions, etc.

---

5 $B;B'$ should not be taken as the union of background conditions. Some conditions in B' may be inconsistent with some in B.

An implementation of the constraints and the function N

Here, we discuss a simple implementation that derive constraints from an on-line manually-built thesaurus known as WordNet [6]. In this way, correct relationships are provided. WordNet represents polysemy information, from which we derive the background conditions and a formulation of N.

Consider the WordNet synonyms of the terms "mortal":

> *Sense 1*
> person, individual, someone, human, soul

Since the term "mortal" has one sense, the fact that "mortal" occurs in a document means that the terms "individual", "someone", etc., are also implicit in the document. Therefore, an unconditional constraint can be derived from this entry.

In WordNet, the synonyms of the term "computer" are the following:

> *Sense 1*
> data processor, electronic computer, information processing system
>
> *Sense 2*
> calculator, reckoner, figurer, estimator

Since the term "computer" is polysemic, its synonyms depends on its sense. Two conditional constraints can be derived: one with sense 1 of "computer", and the other with sense 2. The first constraint relates "computer" to, for instance, "data processor", whereas, the second constraint relates "computer" to, for instance, "estimator". The background conditions of the two constraints are "the term computer has sense 1" and "the term computer has sense 2", respectively. If the term "computer" appears in a document with senses 1 or 2, then the term is ambiguous. The two constraints can be used, but they both lead to an uncertain flow.

We use the second definition (I2) of the satisfaction of the background conditions to formulate the function N. If the term "computer" appears in a document with sense 1, and in another document with senses 1 or 2, then there are more alternative transformations that can originate from the situation modelling the second document than there is from the situation modelling the first document. Let s and s' be the situations that represent the initial representation of the first document and the second document, respectively. Let B the background conditions associated to the first above constraint (sense 1 of "computer"). In s, the number of alternative transformations with respect to B is none, whereas in s' it is 1. This can be captured by the following formulation of N:

$$N(s,B) = \frac{K}{n_{s,B} + 1}$$

where K is some constant and $n_{s,B}$ is the number of alternative transformations from a situation s given the background conditions B. In the above example, we have:

$$N(s,B) = 1 \text{ and } N(s',B) = 1/2$$

Although this is not the most accurate quantitative representation of the uncertainty of a flow, it is a neutral one if no other additional information is available (e.g., knowing that a sense of a word is more probable to occur in a document, etc., ..). From there, it is possible to define N for sequential and parallel transformations. Two possible definitions are (s' is the intermediary situation in the sequential transformation and $B \cap B'$ is the intersection of background conditions):

$$N(s, B; B') = N(s, B) * N(s', B')$$

$$N(s, B \| B') = N(s, B) + N(s, B') - N(s, B \cap B')$$

These definitions reflect that the more transformations there are, lower the values of N, and the more parallel transformations, the higher the values of N. Other formulations of N may be considered, for example, to include dependency between background conditions.

Conclusion

We have developed a logic-based model of an IR system that aims to capture information and its flow. The model is based on the Transformation Principle and uses Situation Theory so that a transformation corresponds to the flow of information. The flow was represented through the constraints which could be certain and uncertain. The uncertainty of the latter constraints were modelled by background conditions, which constitute qualitative representations of the uncertainty. We show that a quantitative representation of the uncertainty, which is necessary to obtain a ranking output of documents, can be derived from the background conditions. We also show a implementation based on an available on-line thesaurus.

Our approach was simple, but demonstrates that the conversion from quantitative to qualitative representation is achievable. This is important because we want our model to capture the flow of information, <u>and</u> to deliver a ranking output of the documents based on the uncertainty of the flow. A more thorough study of the background conditions should be done to understand better the way they act and inter-act. From there, a more accurate quantitative representation of the uncertainty might be derived.

Reference

1 - Barwise, J., 1989. *The Situation in Logic*, CSLI Lecture Notes 17.

2 - Chiaramella Y. and J.P. Chevallet J.P., 1992. About Retrieval Models and Logic, *The Computer Journal*, 35, 233-242.

3 - Devlin, K.J., 1991. *Logic and Information*. Cambridge University Press.

4 - Dretske, F., 1981. *Knowledge and flow of Information*. Bradford Books, MIT Press.

5 - Lalmas, M.and van Rijsbergen, C.J., 1993. A model of an information retrieval system based on situation theory and Dempster-Shafer theory of evidence, In V.S. Alagar, S. Berger, and F. Dong (eds)*Incompleteness and Uncertainty in Information Systems*, Concordia University, Montreal, pp 62-67.

6 - Miller G.A. (ed), 1990. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, Vol. 3(4).

7 - Saffioti A., 1987. An AI view of the treatment of uncertainty, In *The Knowledge Engineering Review*, Vol. 2 (2), pp 75-97.

8 - van Rijsbergen, C.J., 1986. A non-classical logic for Information Retrieval. *The Computer Journal.* 29, pp 481-485.