Modeling Information Retrieval with Probabilistic Argumentation Systems

Justin Picard

Institut interfaculaire d'informatique
University of Neuchâtel
Pierre-à-Mazel 7
2000 Neuchâtel
Switzerland
Justin.Picard@seco.unine.ch

Rolf Haenni

Institut d'informatique
University of Fribourg
rue Faucigny 2
1700 Fribourg
Switerland

Abstract

Probabilistic Argumentation Systems (PAS) are a technique for representing uncertainty both symbolically and numerically. It is shown that this technique, which combines symbolic logic and probability, can be used as a general model of information retrieval. PAS provide a dual (symbolic and numerical) interpretation of the logical uncertainty principle, and are a flexible model for integrating various sources of information about query or document contents.

1 Introduction

Empirical results have shown that performances of a retrieval system may be improved by proper integration of multiple query representations [TC91,BKFS95] and multiple document representations [RC95], relations between words [RY79, CvR95] and relations between documents [FNL88, Sav95]. These results suggest a new direction in designing IR systems : multiplying the sources of information should compensate partly for their fundamental uncertainty. As stated in the famous Principle of Combination [FNL88]:

" Effective integration of more information should lead to better information retrieval ".

Indeed, the " Combining evidence " paradigm is more and more regarded as one of the most promising ways of improving IR performances. Nevertheless, at this time these evidence have not always been optimally exploited, because IR models are sometimes not general enough to model information sources for which they were not initially conceived. Non-classical sources of evidence are complex to model, and the added evidence is rather difficult to quantify. More general formalisms are needed to model and combine the evidence from various kinds of knowledge. The inference network of Turtle [TC91, 92] has proven powerful at modeling various kind of evidence.

A different and very promising approach to IR is the logical paradigm, in which relevance is computationally defined as the degree to which a query/document can be proved, having a document/query as evidence. This approach has led to very interesting theoretical results, notably by clarifying in a logical way the concept of relevance, and by providing a guideline for IR models. Also it has been argued that logic can represent the flow of information, fundamental to IR [LB96, vRL96]. Logical models certainly have a very promising future in IR [CC92, LB96]. They are one of the most adequate formalisms for representing multimedia IR, which implies a very general and complex structure of documents [CMF96, Lal97].

While in the Combining evidence approach, retrieval is done using multiple sources of evidence about document and query contents, in the logical approach it is done by transforming the initial information using a certain body of knowledge. Nevertheless, these approaches are more complementary than antagonist : in both cases relevance is evaluated by drawing inference chains between documents and the query, and computing the overall degree of certainty of these chains.

Until now, the most developed implementation of the logical paradigm have used modal logic [Nie89, CvR95]. We propose here to take a rather different technique, namely Probabilistic Argumentation Systems (PAS) [KR96]. PAS are a technique which combines symbolic logic with probability to model uncertain knowledge (facts and rules) both symbolically and numerically. In a model of IR based on PAS, the Combining evidence and logical

paradigms can be unified. This model provides a dual (symbolic and numerical) interpretation of the logical uncertainty principle.

We will first present some theory on PAS. Then we will show how it can serve as a model of IR. In section 4, query expansion will be treated using PAS. A discussion on the potential of PAS for modeling IR will end this paper.

2 An introduction to PAS

Classical logic cannot be used to handle, represent and compute numerical uncertainty: it is restricted to certain facts or rules. Nevertheless, it is one of the simplest as well as one of the most powerful ways to encode knowledge, for the purpose of reasoning (making inferences) from that knowledge. But is representing uncertainty with classical logic really impossible ?

In fact, if we add a certain type of propositional symbols called assumptions to represent uncertainty, we can model uncertain facts and rules, as shown below. Facts and rules are true under the condition that specific assumptions are true. The table 1 shows how uncertainty can be represented with classical logic using assumptions.

Type of knowledge	Logical representation	Natural language equivalence
A fact	P_1	" P_1 is true "
A rule	$P_1 \rightarrow P_2$	" P_1 entails P_2 "
An uncertain fact	$a_1 \rightarrow P_1$	" if assumption a_1 is true, then P_1 is true "
An uncertain rule	$a_2 \rightarrow (P_1 \rightarrow P_2)$ $\Leftrightarrow P_1 \wedge a_2 \rightarrow P_2$	" if assumption a_2 is true, then P_1 entails P_2"

Table 1 : Representing uncertainty with classical logic using assumptions

A triple (P,A,Σ), where $P=\{P_1,...,P_N\}$ is the set of propositions representing the N variables of interest, $A=\{a_1,...,a_M\}$ the set of M propositions called assumptions used for representing the uncertainty, and $\Sigma=\{\xi_1,...,\xi_R\}$ a set of facts and rules on literals from A and P, is called a Propositional Argumentation System [Hae96, Hae97]. A Propositional Argumentation System can represent uncertainty symbolically, which is useful to explain decisions taken and renders the inference process transparent. In this text, we will take capital letters for propositions and small letters for assumptions.

Arguments supporting or discounting certain hypotheses are derived from the knowledge base Σ. A hypothesis h is any logical formula with symbols in $A \cup P$. An argument in favor (or against) h is a conjunction of literals of assumptions for which h becomes true (or false). Then the hypothesis h is said to be supported (or discarded) by the argument. The support of h is defined as the disjunction of all minimal arguments supporting h, and is denoted sp(h).

Example 1 : Suppose we have a set of variables of interest $P=\{P_1, P_2\}$ the uncertainty being represented by a set of assumptions $A=\{a_1,a_2,a_3\}$ and a set of rules $\Sigma=\{\xi_1: a_1 \rightarrow P_1, \xi_2: a_2 \rightarrow P_2, \xi_3: P_2 \wedge a_3 \rightarrow P_1\}$, and we wish to test the hypothesis P_1 (P_1 true). We have two arguments in favor of P_1: a_1 and $a_2 \wedge a_3$. The support of P_1 is then $sp(P_1)=(a_1 \vee (a_2 \wedge a_3))$.

We may also be interested in finding evidence against an hypothesis h, or reasons to doubt about h. The doubt of h is defined as the disjunction of all arguments supporting ~h and not supporting h, and is denoted db(h). Alternatively, the less reasons we have to doubt about h, the more plausible it seems. The plausibility of h is defined as pl(h)=~db(h). When both an hypothesis and its negation are supported, there is a contradiction in the knowledge base. The support of the contradiction is the disjunction of all arguments which, if true, entails the contradiction.

Example 2 : To illustrate these new concepts, take example 1 and add the rule $\xi_4: a_4 \rightarrow \sim P_1$. Obviously a_4 is an argument against p_1. The support of P_1 is now $sp(p_1)=(a_1 \vee (a_2 \wedge a_3)) \wedge \sim a_4$. The doubt of P_1 is $db(P_1)= a_4 \wedge \sim(a_1 \vee (a_2 \wedge a_3))$, and the plausibility is $pl(P_1)=\sim db(P_1)=\sim(\sim(a_1 \vee (a_2 \wedge a_3)) \wedge a_4)= (a_1 \vee (a_2 \wedge a_3)) \vee \sim a_4$. Since it is not possible for P_1 and $\sim P_1$ to be true at the same time, there is a contradiction in the knowledge base. The support of the contradiction is $sp(\perp)=(a_1 \vee (a_2 \wedge a_3)) \wedge a_4$

With Propositional Argumentation Systems, the reasoning process is fully described but the uncertainty is only represented symbolically, not assessed. To assess uncertainty, we need to assign probabilities to assumptions, e.g. $p(a_1)=x_1$, $p(a_2)=x_2$ etc. Assumptions are probabilistically independent, i.e. $p(a_1 \wedge a_2)=p(a_1).p(a_2)$. Adding the set X of probabilities of assumptions to the triple (A,P,Σ), we obtain a Probabilistic Argumentation System (PAS). From

the support of a hypothesis and the probabilities assigned to assumptions, we can compute a numerical degree of support dsp(h) of the hypothesis, but we need first to put its symbolic support in disjoint form. Different algorithms have been developed for transforming a logical expression in disjoint form, see [Hei89, Mon96, Abr79]. A numerical degree of support is always between 0 and 1, but must not be assimilated with a probability.

Example 3 : In example 1, we want to calculate the numerical support $dsp(P_1)$ of hypothesis P_1 from the symbolic support $sp(P_1)=a_1 \vee (a_2 \wedge a_3)$. We have $p(a_1)=0.5$, $p(a_2)=0.6$, $p(a_3)=0.3$. $sp(P_1)$ must be first put in disjoint form : $sp(P_1)= a_1 \vee (a_2 \wedge a_3)= a_1 \vee (a_2 \wedge a_3 \wedge \neg a_1)$. We then have $dsp(P_1) = 0.5+0.6*0.3*(1-0.5) = 0.59$.

In the case of a partly contradictory knowledge base, the degree of support is normalized by taking into account the support of the contradiction in calculating the support of a hypothesis.

Example 4 : Take example 2 with $p(a_4)=0.2$ We have $sp(\perp)=(a_1 \vee (a_2 \wedge a_3)) \wedge a_4=(a_1 \vee (a_2 \wedge a_3 \wedge \sim a_1)) \wedge a_4$, then $dsp(\perp)=0.59*0.2=0.118$. $sp(p_1)=(a_1 \vee (a_2 \wedge a_3)) \wedge \sim a_4=(a_1 \vee (a_2 \wedge a_3 \wedge \sim a_1)) \wedge \sim a_4$. Then the normalized numerical degree of support of P_1 is :

$$dsp(P_1)=\frac{dsp(a_1 \cup (a_2 \mid a_3 \mid \sim a_1) \mid \sim a_4)}{1-dsp(\perp)}=\frac{059*08}{1-0118} \cup 0535$$

This survey of the theory is sufficient for understanding its application to IR. The reader may have noticed a similarity between Assumption Truth Maintenance Systems (ATMS) [dKl86] and PAS. In fact PAS are an extension of ATMS. While ATMS are limited to Horn clauses, PAS can handle any kind of clauses, for instance, rules like $P_1 \wedge \sim P_2 \wedge a \rightarrow \sim P_3$. Also, it can be shown that PAS is a concrete model of a general theory of evidence [Koh95].

3 Modeling IR with PAS

3.1 The logical approach and PAS

Comparison of different retrieval models of IR has led van Rijsbergen to argue that IR is a form of uncertain inference, each model having its own way to assess uncertainty [vR86]. This led to the logical approach to IR (for a survey, see [Lal96]). This approach, as reformulated in [CC92], states that:

1. In order to be relevant to a query Q, a document D must logically imply Q : $D \rightarrow Q$.
2. Since information is by nature uncertain in IR, the truth of this implication cannot be established with certainty, and we can only measure a degree of certainty $P(D \rightarrow Q)$.
3. This degree of certainty is evaluated through the bias of a logic, following a general uncertainty principle, which in the present case can be enunciated as follows :

   Given a query Q and a document D, a measure of the certainty of $D \rightarrow Q$ is given by the minimal amount of information that must be added to D in order that $D \rightarrow Q$.

Nie [89] proposed an extension to this approach, to take into account two different aspects of relevance. For some users, a document is relevant if it covers all aspects of the query, and relevance is interpreted as exhaustivity ($D \rightarrow Q$). For others, a document must be specific to the query in order to be relevant : relevance is rather interpreted as exclusivity ($Q \rightarrow D$). This leads to a more general evaluation of relevance, composed of these two properties of relevance : $R(D,Q)=F[P(D \rightarrow Q),P'(Q \rightarrow D)]$. We believe that in practice, any interpretation of relevance (($D \rightarrow Q$) or ($Q \rightarrow D$)) can be used depending on the specific IR problem to be modeled with logic. In some cases, rules for inferring the query are better suited while in others, rules for inferring documents are more adequate.

With PAS, the evaluation of $P(D \rightarrow Q)$ (or $P'(Q \rightarrow D)$) can take either a symbolic or a numerical form. For a detailed explanation of the inference process, we take the logical interpretation in which $P(D \rightarrow Q)$ is evaluated by the support of Q once D (and no other document) is considered true (we add D to the knowledge base), which is denoted $sp_D(Q)$. The minimal amount of information that must be added to D for $D \rightarrow Q$ is expressed in the form of a set of arguments (a logical formula containing assumptions). For a numerical evaluation, we put the arguments in disjoint form and assign probabilities to assumptions, and thus compute a numerical degree of support denoted $dsp_D(Q)$. Equivalently, $P'(Q \rightarrow D)$ is evaluated by the support of D once Q is set to true, denoted respectively $sp_Q(D)$ and $dsp_Q(D)$ for symbolic and numerical support.

This dual evaluation of uncertainty is rather new in IR. Nevertheless it corresponds to the profound nature of IR, which can be approached both logically and probabilistically. Different authors have outlined that retrieval can

be modeled by logical inference [vR86, 89, CC92, Lal96], and it has been shown that many retrieval models can be reformulated with logic [vR86, Nie89]. But retrieval may also be viewed as an evidential reasoning process [TC90] based on multiple sources of evidence, where the probabilistic nature of information is fundamental. With PAS the role of logic and probability are clearly distinguished. Logic is used to represent uncertainty and drawing inferences, while probability is used to evaluate uncertainty. An interesting aspect of PAS is that the inference process is completely explicit : the retrieval system can " explain " why a document is judged relevant (or not).

An important issue has not been addressed : how are the probabilities of assumptions assessed ? Since an assumption represents the uncertainty of a specific rule/fact, the probability of this assumption should be equal to the probability that this rule/fact is true. For example, having a rule $P_1 \wedge a_1 \rightarrow P_2$, the assumption a_1 should be evaluated by $p(a_1)=p(P_2|P_1)$. Of course probability estimation is one of the fundamental problems of IR [TC97], and this is one of the main problem we face when modeling IR with PAS. We will discuss that issue in Section 5.

3.2 An example

To see how a retrieval system can be modeled with PAS, consider the following example. A document D is represented by terms T_1, T_2 and T_3, and a query Q by terms T_1 and T_3. The set of variables of interest is P={D, Q, T_1, T_2, T_3}.

Suppose we take the D→Q interpretation of relevance, evaluated by $sp_D(Q)$. We need rules for inferring Q once D is set to true. The set of rules is : {D, $D \wedge a_1 \rightarrow T_1$, $D \wedge a_2 \rightarrow T_2$, $D \wedge a_3 \rightarrow T_3$, $T_1 \wedge b_1 \rightarrow Q$, $T_2 \wedge b_2 \rightarrow Q$} and the set A={$a_1$,$a_2$,$a_3$,$b_1$,$b_3$} represents the uncertainty. Clearly the support of Q is $sp_D(Q)=(a_1 \wedge b_1) \vee (a_3 \wedge b_3)$ or $sp_D(Q)=(a_1 \wedge b_1) \vee ((a_3 \wedge b_3) \wedge \neg(a_1 \wedge b_1))$ in disjoint form.

If we take the Q→D interpretation of relevance, we need rules for inferring D . The set of rules is reversed: {Q, $Q \wedge c_1 \rightarrow T_1$, $Q \wedge c_3 \rightarrow T_3$, $T_1 \wedge d_1 \rightarrow D$, $T_2 \wedge d_2 \rightarrow D$, $T_3 \wedge d_3 \rightarrow D$,}. Clearly the support of D is $sp_Q(D)=(c_1 \wedge d_1) \vee (c_3 \vee d_3)$

The figure 1 shows the two PAS corresponding to the two interpretations of relevance.

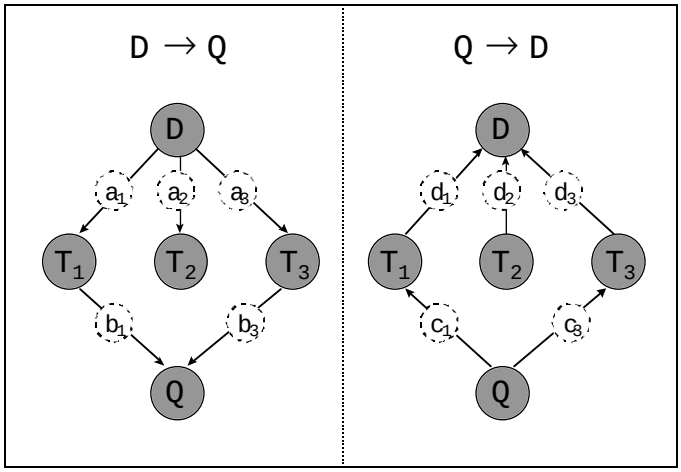

Figure 1 : Representation of the PAS for the D→Q and Q→D interpretations of relevance

Until now, we have only a symbolical model of IR, and we need to assess assumptions to have a complete model which can rank documents relatively to a certain query. We will make the two very simple assumptions :

- if a document or a query contains a term there is a 0.5 probability that the document/query entails that term
- if a document or a query contains N term, there is a 1/N probability that the term entails the document/query

We then have $p(a_1)=p(a_2)=p(a_3)=0.5$, $p(b_1)=p(b_2)=0.5$, $p(c_1)=p(c_2)=0.5$, $p(d_1)=p(d_2)=p(d_3)=0.33$. The two degrees of support are :

$dsp_D(Q)=(0.5*0.5)+(0.5*0.5*(1-0.5*0.5))=0.4375$

$dsp_Q(D)=(0.5*0.33)+(0.5*0.33*(1-0.5*0.33)=0.3027$

Logically the measure of exhaustivity is superior to the measure of specificity, since all query terms are document terms, while the converse is not true.

3.3 A model of IR based on PAS

To design a PAS , we must first choose our variables of interest, represented by a set P of propositions. We will consider only one document for sake of simplicity. The set P contains :

- D which represents the original document.
- $D_1,...,D_R$, which represent the R different document representations.
- $T_1,...,T_S$, which represent the indexing terms (stems, phrases) used to represent documents and queries.
- Q which represents the original query.
- $Q_1,...,Q_U$, which represent the original query and the different query representations.

For the D→Q interpretation of relevance, the kind of rules used are :

Kind of rule	Interpretation
$D \wedge d_i \rightarrow D_i$	D entails document representation D_i if assumption d_i is true
$D_i \wedge t_j \rightarrow T_j$,	D_i entails term T_j with assumption t_j
$T_j \wedge l_{jk} \rightarrow T_k$,	T_j entails T_k with assumption l_{jk} (T_j and T_k are related terms)
$T_j \wedge r_{jm} \rightarrow Q_m$	T_j entails query representation Q_m with assumption r_{jm}
$Q_m \wedge q_m \rightarrow Q$	Q_m entails the original query Q with assumption q_m

Table 2 : Rules in the D→Q interpretation of relevance

Figure 2 shows an example. Dashed circles represent assumptions. The document D has 3 different document representations. There are 4 indexing terms of which two are related (T_3 to T_4). Query Q has 2 different representations. The support of Q can be found by tracing all paths (inference chains) from D to Q.

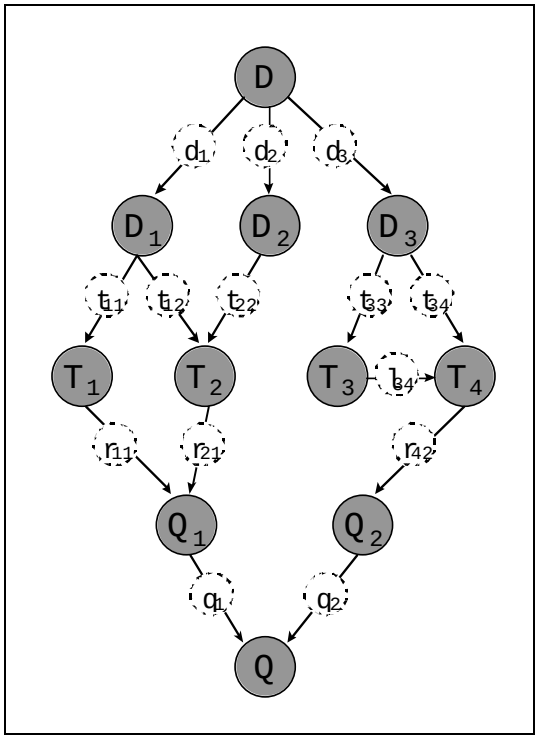

Figure 2 : Example of a PAS for modeling IR

4 PAS for query expansion

While there is a great hope that uncertain logics may help building powerful IR models, they are often considered too

computationally expensive to be used in large scale IR (but see [CRSvR96]). Another handicap of logical models is that commercial retrieval systems are well established and it is costly to change them completely. But uncertain logics can be used to help solving specific IR problems, working at a precise stage of the retrieval process, without changing fundamentally the retrieval system. Also, a general model of IR should not be built solely on theoretical considerations : combining theoretical investigation with experiments on practical problems of IR should lead to a better understanding of the inference processes. Here we investigate the use of PAS for improving the query representation using positive and negative relationships between terms.

4.1 Modeling query expansion

Our modeling of query expansion is based on the Q→D interpretation of relevance : starting from the query, we try to infer the document. To design a PAS for query expansion, the variables of interest are the initial query Q, document D and all the representation terms. These variables of interest are represented by the set of propositions $P=\{Q, D, T_1,...T_N\}$. The retrieval system computes a score on the terms representing the query, which must be converted into a probability.

Assumption a_i represents the uncertainty on term T_i for representing the query. For all T_i representing the query :

$$Q \wedge a_i \rightarrow T_i$$

Since in this model the query is set to true, we have the equivalent rule :

$$a_i \rightarrow T_i$$

A link between terms T_i and T_j is interpreted as information that the presence of T_i is evidence for the presence of T_j. The uncertainty of this information is represented by an assumption l_{ij} :

$$T_i \wedge l_{ij} \rightarrow T_j$$

In a similar way, negative evidence (T_i is evidence for absence of T_k) is modeled this way :

$$T_i \wedge l_{ik} \rightarrow \sim T_k$$

Note that multiple relationships between words T_1 and T_2 when combining different body of knowledge (for example when combining relationships from thesauri, statistical co-occurrence and pseudo-classification) can be modeled by different assumptions:

$$T_1 \wedge l_{12} \rightarrow T_2,\ T_1 \wedge l'_{12} \rightarrow T_2,\ T_1 \wedge l''_{12} \rightarrow \neg T_2.$$

Assumption b_i represents the uncertainty on term T_i for representing the document. For all T_i representing the document :

$$T_i \wedge b_i \rightarrow D$$

The main purpose of using PAS here is to provide a theoretical framework for making inferences using term relationships. If the PAS is used solely for expanding the query in a well-established IR system based for example on the vector-space model, a simple matching can be done once the numerical support of each query term is computed. The rules for inferring document D are necessary only if the whole system is based on the PAS framework.

We are presently investigating ways to assess the probabilities of link assumptions. Nie and Brisebois [96] propose a very interesting way to learn the strength (between 0 and 1) of thesauri relationships using previous relevance judgments, within a fuzzy modal logic framework. On the CACM collection with Wordnet™ thesaurus and a set of 50 queries for training, they find approximately a strenght of 0.1 for synonymy relationships, 0.3 for holonoymy and 0.85 for meronymy. Of course these values are thesaurus, collection, test queries and system dependent. We intend to adapt their method to PAS.

Statistical co-occurrence information has not always shown useful to IR : Peat and Willet [91] give an explanation to that paradox. Second-order co-occurrence is more reliable : with that technique, a term is represented by a vector of all terms with which it occurs in a certain context. The context can be for example a sentence, a paragraph, a document or a sliding window. Then a measure of similarity (typically a cosine measure) can be computed for every term pairs. With that method, two synonyms like " color " and " colour " (which rarely occur in the same document) should have a high measure of similarity since they are usually found with the same words. The complexity of computing second-order co-occurrence is $O(N^3)$ with the number of different terms N, but Schütze and Pedersen [97] show how to reduce it to $O(N^2)$ using singular value decomposition.

We are still investigating ways for converting a similarity measure to the probability of a link assumption, but the general idea is as follows : for a specific word, we compute its similarity with all the other words, and find the average similarity for this word. This word is then considered as positive/negative evidence for the presence of words with a similarity measure higher/lower than the average similarity.

In our preliminary investigation on the CISI collection we have computed the cosine similarity measure of the

word ‘information’ with all words found more than 10 times in the collection. The average similarity measure is 0.35. Two highly correlated terms are ‘data’(0.61) and ‘retrieval’(0.76), while the word ‘game’ has a similarity of 0.18. Assume we compute the probability of a link assumption as the difference between the similarity measure and the average similarity measure, we find that ‘information’ entails ‘retrieval’ with a probability of 0.41 :

$$\text{‘information’} \wedge l_{12} \rightarrow \text{’retrieval’} \quad \text{with } p(l_{12})=0.76-0.35=0.41$$

The probability of link assumption is not symmetric : ‘retrieval’ has an average similarity with the other words of only 0.24. It is then stronger evidence for the presence of ‘information’ than the converse :

$$\text{‘retrieval’} \wedge l_{21} \rightarrow \text{’information’} \quad \text{with } p(l_{21})=0.76-0.24=0.52$$

Since word ‘game’ has a low similarity measure with ‘information’, the latter should be considered as negative evidence for the former :

$$\text{‘information’} \wedge l_{14} \rightarrow \sim\text{’game’} \quad \text{with } p(l_{14})=-(0.18-0.35)=0.17$$

Negative evidence should help discount “ noisy ” words which are added from a manual thesauri but should not be related in the context of the query. There is still a lot of investigation for finding a proper way of computing link assumptions, for building an efficient inference engine to compute the support of each term, and for reducing the amount of computation required.

4.2 An example

A query is represented by terms T_1, T_2 and T_3. As in section 3.2, there is a prior support of 0.5 on these terms. Assume that terms T_1, T_2 and T_3 are respectively linked T_4 and T_5 T_6 and T_7, T_8 and T_9. by thesaurus relationships. For sake of simplicity, we do not consider co-occurrence relationships, which can be negative. Also, we do not consider for this example words that would be linked with T_4 to T_9, but we consider “ inside ” links between T_1 to T_9. Assume there are three types of relationships, with probability 0.6, 0.4 and 0.2.

The set of rules is :

$a_1 \rightarrow T_1$	$T_1 \wedge l_{14} \rightarrow T_4$	$T_2 \wedge l_{27} \rightarrow T_7$	$T_2 \wedge l_{28} \rightarrow T_8$	$T_8 \wedge l_{87} \rightarrow T_7$
$a_2 \rightarrow T_1$	$T_1 \wedge l_{15} \rightarrow T_5$	$T_3 \wedge l_{38} \rightarrow T_8$	$T_3 \wedge l_{36} \rightarrow T_6$	
$a_3 \rightarrow T_3$	$T_2 \wedge l_{26} \rightarrow T_6$	$T_3 \wedge l_{39} \rightarrow T_9$	$T_6 \wedge l_{69} \rightarrow T_9$	

The probabilities of assumptions are :

$p(a_1)=p(a_2)=p(a_3)=0.5$

$p(l_{14})=p(l_{26})=p(l_{38})= p(l_{69})=p(l_{87})=0.6$

$p(l_{15})=p(l_{27})=p(l_{39})=0.4$

$p(l_{36})=p(l_{28})=0.2$

The PAS is represented on the figure below.

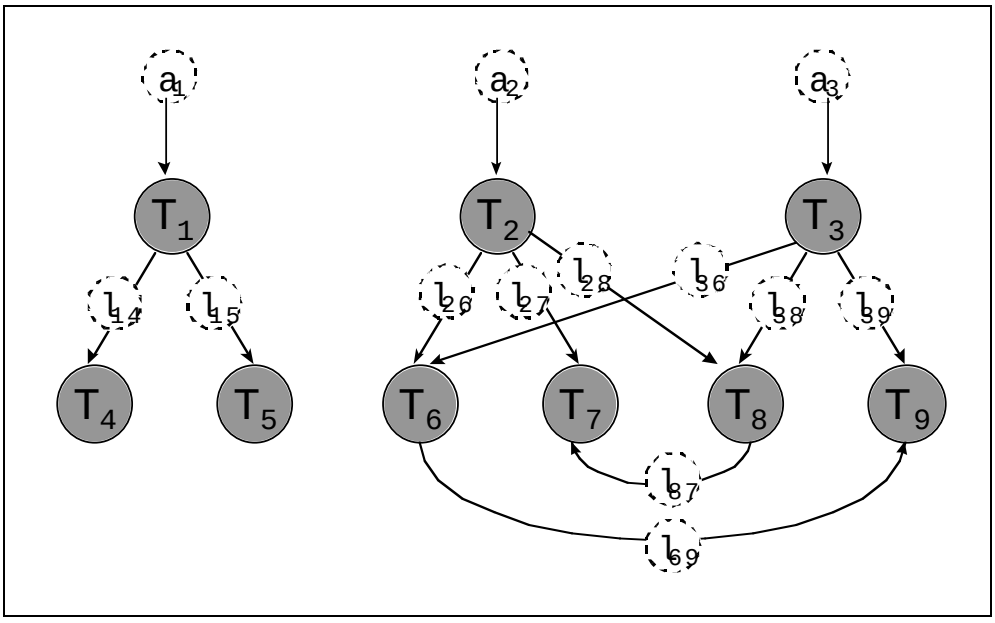

Figure 3 : Example of a PAS for query expansion

The symbolic support of each term can be easily deduced on figure 3. It is :

$sp(T_1)=a_1$

$sp(T_4)=a_2$

$sp(T_3)=a_3$

$sp(T_4)=a_1 \wedge l_{14}$

$sp(T_5)=a_1 \wedge l_{15}$

$sp(T_6)=(a_2 \wedge l_{26}) \vee (a_3 \wedge l_{36})$

$sp(T_7)=(a_2 \wedge l_{27}) \vee (a_2 \wedge l_{28} \wedge l_{87}) \vee (a_3 \wedge l_{38} \wedge l_{87})$

$sp(T_8)=(a_2 \wedge l_{28}) \vee (a_3 \wedge l_{38})$

$sp(T_9)=(a_3 \wedge l_{39}) \vee (a_2 \wedge l_{26} \wedge l_{69}) \vee (a_3 \wedge l_{36} \wedge l_{69})$

We must put the symbolic support of T_6 to T_9 in disjoint form to compute their degree of support. With Heidtmann's algorithm :

$sp(T_6)=(a_2 \wedge l_{26}) \vee ((a_3 \wedge l_{36}) \wedge \sim(a_2 \wedge l_{26}))$

$sp(T_7)=(a_2 \wedge l_{27}) \vee (a_2 \wedge l_{28} \wedge l_{87} \wedge \sim l_{27}) \vee (a_3 \wedge l_{38} \wedge l_{87} \wedge \sim a_2) \vee (l_{38} \wedge l_{87} \wedge a_2 \wedge a_3 \wedge \sim(l_{28} \wedge l_{27}))$

$sp(T_8)=(a_2 \wedge l_{28}) \vee ((a_3 \wedge l_{38}) \wedge \sim(a_2 \wedge l_{28}))$

$sp(T_9)=(a_3 \wedge l_{39}) \vee (a_2 \wedge l_{26} \wedge l_{69} \wedge \sim(a_3 \wedge l_{39})) \vee ((a_3 \wedge l_{36} \wedge l_{69}) \wedge (\sim l_{39}) \wedge \sim(a_2 \wedge l_{26}))$

The degrees of support are :

$dsp(T_1)=0.5$

$dsp(T_2)=0.5$

$dsp(T_3)=0.5$

$dsp(T_4)=0.5*0.6=0.3$

$dsp(T_5)=0.5*0.6=0.3$

$dsp(T_6)=(0.5*0.6)+(0.5*0.2)*(1-0.5*0.6)=0.37$

$dsp(T_7)=(0.5*0.4)+(0.5*0.2*0.6*(1-0.4))+(0.5*0.6*0.6*(1-0.5))+(0.6*0.6*0.5*0.5*(1-(0.2*0.4)))=0.3827$

$dsp(T_8)=(0.5*0.2)+(0.5*0.6)*(1-0.5*0.2)=0.37$

$dsp(T_9)=(0.5*0.4)+(0.5*0.6*0.6*(1-0.5*0.4))+(0.5*0.2*0.6*(1-0.4)*(1-0.5*0.6)=0.3692$

In summary, we started from a query Q={T_1 :0.5,T_2 :0.5,T_3 :0.5}. The expanded query is :

Q'={T_1 :0.5, T_2 :0.5, T_3 :0.5, T_4 :0.3, T_5 :0.3, T_6 :0.37, T_7 :0.3827, T_8 :0.37, T_9 :0.3692}

T_4 to T_9 are the added terms in the expanded query. Computing T_4 and T_5's probability is straightforward. The support of terms T_6 to T_9 illustrate how evidence is propagated with PAS : a piece of evidence is never counted twice. In practice, making a query expansion with PAS will probably entail much more related terms and complex computations.

5 Discussion

More general formalisms are needed to model and combine the evidence from various kinds of knowledge. The inference network model of Turtle [TC90, 91] has emerged as powerful for modeling various kind of uncertain knowledge. In this approach, probabilistic causal relationships between variables are combined in order to estimate the probability that a document meets a user's information need. Different formulation of a query, multiple document representation, etc. are very naturally modeled within this model. Nevertheless, some types of knowledge are not easily modeled : special attention must be put to prevent cycles in the network (evidence would be propagated indefinitely), multiple relationships between variables must be summarized in one link matrix. Also, this approach is purely numerical, and the system cannot easily explain its decisions. PAS could then be considered as an alternative solution to the inference network model, for solving these specific problems.

While the need of a logic as a formal model of IR has been strongly justified by theoretical arguments, it is not always obvious how a logic would solve practical IR problems. We think that logic can serve for modeling specific knowledge, which cannot be adequately addressed with classical methods. Non-classical sources of evidence, which seem to be a very promising way to improve IR performances, need special methods to be adequately modeled.

In further research, we will concentrate on assessing probabilities of assumptions, in order to make different practical applications of PAS to IR. It is of course possible to give ad-hoc or " well-suited " values to probabilities, but our intention is to base probability estimates on strong theoretical arguments. In another paper [Pic98], the problem of modeling and combining evidence provided by document relationships is tackled with PAS. It is shown how prior probabilities on document's relevance is assessed by a logistic regression using the rank. It is also shown how to assess the probability that a document is relevant if it is linked to a document known to be relevant. A practical implementation was made on the CACM collection, with satisfying results.

Our intention now is to develop more thoroughly the model for modeling document relationships, and to make a practical implementation of PAS for query expansion. Then, we may think of implementing a complete IR system with PAS.

6 Acknowledgments

The authors wish to thank Jacques Savoy for his useful comments and suggestions. This research was supported by the SNSF (Swiss National Scientific Foundation) under grants 21-49427.95.

7 Reference List

[Abr79] J.A. Abraham. An improved algorithm for network reliability. *IEEE Transactions on Reliability*, 28, 58-61.

[BKFS95] N.J. Belkin, P. Kantor, E.A. Fox and J.A. Shaw. Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing and Management*, 31(3), 431-448, 1995.

[BM96] R. Bertschy and P.A. Monney. A Generalization of the Algorithm of Heidtmann to Non-Monotone Formulas. *Journal of Computational and Applied Mathematics*, 76, 55--76. 1996.

[CC92] Y. Chiamarella and J.P. Chevallet. About Retrieval Models and Logic. *The Computer Journal*. *5(3), 233-242, 1992.

[CMF96] Y. Chiamarella, P. Mulhem and F. Fourel. A model for multimedia Information Retrieval. Technical Report, Basic Research Action FERMI 8134, 1996.

[CvR95] F. Crestani and C.J. van Rijsbergen. Probability Kinematics in Information Retrieval. In *Proceedings of ACM SIGIR*, 291-299, Seattle, WA, USA, 1995.

[CRSvR96] F. Crestani, I. Ruthven, M. Sanderson and C.J. van Rijsbergen. The troubles with using a logical

model of IR on a large collection of documents. *TREC-4*, 509-526. 1996.

[DTFLH90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407, 1990.

[dKl86] J. de Kleer. An Assumption-Based TMS. *Artificial Intelligence*, 28, 127-162, 1986.

[FNL88] E.A. Fox, G.L. Nunn and W.C. Lee. Coefficients for Combining Concept Classes in a Collection, *Communication of the ACM*, 1988.

[Hae96] R. Haenni. *Propositional Argumentation Systems and Symbolic Evidence Theory*. Ph.D. Thesis. Institut für Informatik, Universität Freiburg. 1996.

[Hae97] R. Haenni. Modeling Uncertainty in Propositional Assumption-Based Systems. S. Parson (eds.), *Uncertainty in Information Systems*. 1997.

[Hei89] K.D. Heidtmann. Smaller sums of disjoint products by subproduct inversion. *IEEE Transactions on Reliability*, 38(3), 305-311.

[Koh95] J. Kohlas. Mathematical Foundations of Evidence Theory. *Pages 31--64 of:* G. Coletti and D. Dubois and R. Scozzafava (eds.), *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*. Plenum Press. 1995.

[KM95] J. Kohlas and P.A. Monney. *A Mathematical Theory of Hints. An Approach to Dempster-Shafer Theory of Evidence*. Lecture Notes in Economics and Mathematical Systems, vol. 425. Springer-Verlag. 1995.

[KR96] J. Kohlas and R. Haenni. Assumption-Based Reasoning and Probabilistic Argumentation Systems. J. Kohlas and S. Moral (eds.), *Defeasible Reasoning and Uncertainty Management Systems: Algorithms*. Oxford University Press. 1996.

[LB96] M. Lalmas and P.D. Bruza, The use of Logic in Information Retrieval Modeling. Tutorial in *ACM-SIGIR*, Zurich, 1996.

[Lal97] M.Lalmas. Dempster-Shafer's Theory of Evidence Applied to Structured Documents : Modeling Uncertainty. *Proceedings of ACM-SIGIR*, Philadelphia (PA), July 1997, 110-118.

[NB96] J. Nie and M. Brisebois. An inferential approach to information retrieval and its implementation using a manual thesaurus. *Artificial Intelligence Review*, 10, 409-439, 1996.

[Nie89] J. Nie. An information Retrieval Model Based on Modal Logic. *Information Processing and Management*, 25(5), 477-491, 1989.

[Pic98] J. Picard. Modeling and combining evidence provided by document relationships using probability argumentation systems. In *Proceedings of ACM SIGIR*, Australia, 1998. Accepted for publication.

[PW91] H.J. Peat and P. Willet. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 41(4), 288,297, 1991.

[RY79] V.V Raghavan and C.T. Yu. Experiments on the Determination of the Relationships Between Terms. *ACM Transaction on Database Systems*, 4(2), 240-260, 1979.

[RC95] T.B. Rajashekar and W.B. Croft. Combining Automatic and Manual Index Representations in Probabilistic Retrieval. *Journal of the American Society of Information Science*, 46(4), 272-282, 1995.

[vR86] C.J. van Rijsbergen. A non-classical logic for Information Retrieval. *The Computer Journal*, 29(6), 481-485, 1986.

[vR89] C.J. van Rijsbergen. Toward an Information Logic. *Proceedings of ACM-SIGIR* , New York, 77-89, 1989.

[vRL96] C.J. van Rijsbergen and M.Lalmas. Information Calculus for Information Retruieval. *Journal of the American Society of Information Science*, 47(5), 385-398, 1996.

[Sav96] J.Savoy. An extended Vector-Processing Scheme for Searching Information in Hypertext Systems. *IPM*, 32(2), 155-170, 1996.

[SP97] H. Schütze and J.O. Pederesen. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, 33(3), 307-318, 1997.

[TC91] H.R. Turtle et W.B. Croft. Evaluation of an Inference Network-Based Retrieval Model. *ACM Transactions on Information Systems*, 9(2), 187-222, july 1992.

[TC92] H.R. Turtle and W.B. Croft. A comparison of Text Retrieval Models. *The Computer Journal*, 35(3), 279-290, 1992.

[TC97] H.R. Turtle and W.B. Croft. Uncertainty in Information retrieval Systems. In *Uncertainty Management in Information Systems*, A. Motro, P Smets (Ed.), Kluwer, Amsterdam (NL), 1997.