Fuzzy Model Fragment Retrieval

Shen, Qiang; Fu, Xin

Published in:
Proceedings of the 17th International Conference on Fuzzy Systems

Publication date:
2008

Citation for published version (APA):
Shen, Q., & Fu, X. (2008). Fuzzy Model Fragment Retrieval. In *Proceedings of the 17th International Conference on Fuzzy Systems* (pp. 1381-1388) http://hdl.handle.net/2160/596

Fuzzy Model Fragment Retrieval

Xin Fu and Qiang Shen

***Abstract*— Given a set of collected evidence and a knowledge base, Fuzzy Compositional Modelling (FCM) begins by retrieving model fragments which are the most likely to be relevant to the available data. Since FCM often involves imprecise and uncertain information, a match between the available data and the knowledge base cannot in general be done precisely, partial matching may suffice. This paper proposes a more flexible fuzzy model fragment retrieval mechanism to match data items with broader, including possibly subjective information in the knowledge base. It is capable of retrieving those model fragments that can approximately match the collected evidence, when no exact match occurs. The retrieval process and its capability is illustrated by means of an application example.**

I. Introduction

Compositional Modelling (CM) [7], [12] has been developed to synthesize and store plausible scenario spaces in many problem domains with promising results. However, for applications like crime detection and prevention, the notion of vagueness and uncertainty is often involved. Vagueness concerns concepts for which there are no exact definitions, such as *high explosive* materials, *extremist* organizations and *substantial amount* of fibers. When it comes to uncertainty, often, due to lack of knowledge, propositions can not always be stated as true or false. They can only be estimated to which probability/possibility degree they are true or false.

Given a set of collected evidence and a knowledge base, Fuzzy Compositional Modelling (FCM) [8] begins by retrieving model fragments which are the most likely to be relevant to the concepts/predicates involved in the set of collected evidence. As aforementioned, the degree of precision of the available data can be very variable, the collected evidence and model fragments in the knowledge base involve both vague and uncertain information, so that finding a match between them cannot in general be done precisely. The retrieval mechanism developed in this work aims to match specific data items with broader and possibly subjective information in the knowledge base and to pick up those matches exceed a predefined threshold. For example, when a car is observed on a CCTV camera, some identifying information can be collected but this may be insufficient to identify the exact model of the car. Therefore, model fragments which involve similar or more specific features to those of the observed car should also be retrieved.

The proposed mechanism consists of three component approaches: the search component, the match component and the aggregation component. The computation cost of the model fragments retrieval process is highly relevant to the search strategy used as well as the data structure employed to store information. Thus, in building the knowledge base, the atomic concepts/predicates are hierarchically structured in weighted taxonomies and a hash table is employed to establish a link between the atomic concepts/predicates and their related model fragments. Then, two different levels of match, namely sematic matching and fuzzy set matching, are performed by the match component. Finally, the individual fuzzy matching degrees are aggregated in a hierarchical manner to produce the final Retrieval Status Value (RSV) to evaluate the relevance of the candidate model fragments.

Xin Fu (email: xxf06@aber.ac.uk) and Qiang Shen (email: qqs@aber.ac.uk) are with the Department of Computer Science, Aberystwyth University, Aberystwyth, UK.

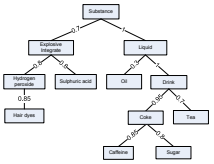

Fig. 1. Portion of a weighted taxonomy

The reminder of this paper is organized as follows. Section II presents the predefined knowledge representation in FCM and a brief overview of fuzzy information retrieval models. Second III proposes a flexible scenario fragments retrieval mechanism. This is followed by an illustrative example in Section IV. Section V concludes this paper and points out future research.

II. Background

A. Knowledge Representation

In order to increase the flexibility of automatically generating plausible scenario spaces, when given pieces of evidence, fuzzy set theory has been applied to the creation of a structured knowledge representation scheme which is capable of storing and managing vague and uncertain data in CM [8]. In particular, a knowledge base consists of the following:

1) Weighted taxonomy: A number of weighted taxonomies are employed to represent a set of concepts or states and their relationships within a given problem domain. There may be many concepts that share structural similarities, therefore, those concepts share something in common or highly relevant are naturally grouped into the same class. Fig. 1 shows an example of (part of) such a weighted taxonomy for the problem domain of counter terrorism.

Note that, not only nodes but also arcs in a given taxonomy can carry semantic information. A weight attached to each arc expresses the degree of relevance between a chid node and its parent. In other words, it indicates to what extent the child node can be classified as an element of the domain of its parent node. The higher weight a concept receives, the more common features it inherits from its parent. These weights are assigned by experts. Note that, since a concept can be classified into different categories, the sum of the

978-1-4244-1819-0/08/$25.00

arc weights of subtype concepts do not have to be 1. Each arc weight can be assigned independently, and subjectively, without taking into account its sibling nodes. Also, the weighted taxonomy is independent of whether the concepts are fuzzy or not. Several taxonomies without the weighting factor which concern with fuzzy concepts such as length, height and level have been defined in [8].

2) Model fragments: For CM, the scenario elements/states and their relationships are modelled as generic and reusable fragments in the knowledge base. A scenario fragment, interchangeably termed model fragment, μ is itself a partial description of a certain scenario, which is defined in a tuple $\langle C^s, C^t, \phi^s, \phi^t, A \rangle$ as follows:

If $\{\phi^s\}$

Assuming $\{A\}$

Then $\{\phi^t\}$

Distribution $\phi^t \ \ \{v_1^s, \ldots, v_{n+s}^s \rightarrow v_1^t, \cdots, v_m^t : q\}$ where

- $C^s(\mu) = \{c_1^s, \cdots, c_k^s\}$ is a set of source-participants, referring to already identified objects/concepts in the partial scenario.
- $C^t(\mu) = \{c_1^t, \cdots, c_q^t\}$ is a set of target-participants, representing new objects/concepts that will be added to the partial scenario description if the model fragment is instantiated.
- $\phi^s(\mu) = \{\phi_1^s, \cdots, \phi_n^s\}$ is a set of relations called structural conditions, whose free variables are elements of C^s. Normally, they are represented by predicates or nested predicates.
- $\phi^t(\mu) = \{\phi_1^t, \cdots, \phi_m^t\}$ is a set of relations called post-conditions, whose free variables are elements of $C^s \bigcup C^t$. Normally, they are represented by predicates or nested predicates.
- $A(\mu) = \{a_1, \cdots, a_s\}$ is a set of assumptions, which may be established to be true or false.
- $v^s = \{v_1^s, \cdots, v_{n+s}^s\}$ are the values of the antecedent predicates and assumptions.
- $v^t = \{v_1^t, \cdots, v_m^t\}$ are the values of the consequent predicates.
- q summaries the probability distributions over the possible assignments of ϕ^t.

The **If** statement describes the required conditions for a partial scenario to become applicable and the **Assuming** statement indicates the reasoning environment. With the purpose of performing hypothetical reasoning, this environment specifies the uncertain events and states which are presumed in a partial scenario description. The **Then** statement describes the consequent when the conditions and presumed assumptions hold. The **Distribution** statement indicates the probability distributions of the consequent predicate or those of their relations. It is worth noting that the likelihood is represented as subjective linguistic probabilities such as *slim chance*, *very likely* and *good chance* [11].

Since such constructed knowledge representation formalism involves both vague and uncertain information, fuzzy predicates are also allowed to appear in model fragments. This implies that a fuzzy matching degree (within $[0, 1]$) will be assigned when performing matching between observed evidence and a given model fragment.

3) Hash table: Once a concept/predicate in weighted taxonomies is required, a hash table is employed to support an efficient look up to retrieve the corresponding model fragments. Compared to other associative array data structures, hash tables are most useful when large numbers of records are to be stored. For example, Fig. 2 shows a hash table that associates the fuzzy predicates *Height* and *Degree_of_fight* with their relevant model fragments in the knowledge base.

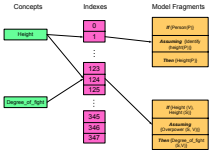

Fig. 2. Example of hash table

It works by assigning an unique index number to each model fragment in the knowledge base, this index number is used to establish a link explicitly between the concept/predicate and the model fragment. In addition, each concept/predicate is attached an index link which stores a sequence of index numbers of those model fragments which involve the defined concept/predicate.

B. Fuzzy Information Retrieval

The goal of an Information Retrieval (IR) system is to automatically retrieve information which satisfies a user's query. Generally, in IR, information is managed at two distinct levels [3]:

- **Representation of information source (typically in the form of documents)**
- **Information request represented through queries**

The *Boolean IR model* [16] is perhaps, one of the most commonly used models in commercial IR systems. Both the documents and queries are represented as sets of index terms. Also, those terms in query are logically connected via boolean operators such as AND, OR and NOT. This model produces an exact answer to indicate a document is relevant or not. However, this crisp behaviour is liable to ignore useful information whereas possibly picking up useless information as a result of selection conditions which are too restrictive. To address this problem, substantial efforts have been devoted in an attempt to develop more flexible IR systems, including systems built on fuzzy set theory [2], [5], [13].

1) Fuzzy extension of document representation: In fuzzy IR systems, a document is typically represented as a fuzzy set of terms - $\{F(d,t)/t\}$ and the definition of the indexing function $F(d,t)$ has drawn much attention to the development of fuzzy IR systems. For instance, a new indexing function has been proposed in [2], which computes the significance of a term in a document by considering the different roles of term occurrences. More precisely, the weight of a term t in a given document d is computed by firstly calculating the weight of t in each subsection individually and then by aggregating the resulting scores using a user-defined function (in order to reflect the varying degrees of significance of different subsections).

Obviously, the incorporation of weighed document representation softens the crisp boolean match, thereby allowing for partial matching. A Retrieved Status Value (RSV) is employed to evaluate the degree of relevance of a document with respect to a given query. In doing so, a ranking of the retrieved documents can be presented in decreasing order of their RSVs. This is more convenient for the user to access the most relevant documents.

2) Fuzzy extension of query representation: Query terms formulated in *Boolean IR models* are being treated equally and connected by boolean operators. Fuzzy extension of such a query representation involves two stages:

The first is to extend the selection criteria (to determine which documents to retrieve) by attaching a weight to each search term in a query. A user can provide a quantitative description of the "weights"of that term in the search documents. The weight in the query can be interpreted as an important weight, as a threshold, or as a description of the "ideal" document. Weights have been initially denoted by numeric values. However, the use of numeric weights requires clear knowledge of their semantics to translate fuzzy concept into a precise numeric value. So in [1], the numeric query weights are replaced by linguistic descriptors which specify the degree of importance of the term. Linguistic weights such as *important* and *fairly important* are formalized.

The second stage is to soften the aggregation process where query evaluation is done firstly for each weighted term and then combining such evaluation according to the query structure. Various aggregation operators can be used to combine single selection criteria to represent more complex query requests. The degree of relevance of each single selection criterion are aggregated together to compute the final RSV of the document with respect to the query. Recently, more and more efforts have been made to define new aggregation operators as a compromise between the AND and OR operators. For instance, the linguistic quantifiers (e.g. at least k) in [2], the Ordered Weighted Averaging (OWA) operators in [17] and the *and possible* operator in [18].

III. The Approach

Given a set of collected evidence, the model fragment retrieval mechanism proposed herein plays the role of retrieving a set of the relevant model fragments ordered according to their RSVs from the knowledge base. This is of course the same in principle as with existing fuzzy IR techniques, but the details of the approach are of its own characteristics. These are described below (using crime detection and prevention as the problem domain).

A. Representation of model fragments

The general form of model fragments has been presented in section II-A.2. Each model fragment consists of three components: antecedent, assumption and consequent. Each component consists of one or more atomic predicates. Since the basic inference methods used in FCM are backward chaining and forward chaining in which only one component per fragment will be examined at a time. These components are of equal importance. However, different search priorities may be assigned to different components with regards to different inference methods. For example, in backward chaining, a priority vector $[0, 0, 1]$ can be assigned, which indicates only the consequent component alone is used to perform the match. Also, in forward chaining, a priority vector $[1, 1, 0]$ may be given.

B. Representation of collected evidence

Given a new or ongoing investigation, an initial set of evidence has been collected by investigators and this collected evidence is entered into the system, triggering the need to generate a space of possible scenarios. The set of collected evidence E can be represented by:

$$E = \{e_1, e_2, \ldots, e_n\} \tag{1}$$

where $e_1, e_2, \ldots, e_n$ are atomic pieces of information that are considered to be observed consequences of a possible crime scenario.

Similar to the weighted IR model, every piece of evidence is attached with a weight in the present work. Such a weight is interpreted as confidence weight which associates a degree of certainty with the collected evidence. For example, if a piece of evidence is a clear CCTV camera observation, it may be assigned a high confidence weight. On the other hand, if a piece of evidence is collected by interviewing a possible witness, this evidence may only receive a relatively low confidence weight. Confidence weights are initially assigned by investigation experts. Such assessment typically reflects the expertise and knowledge of the investigators and is naturally represented in linguistic terms.

C. Outline of model fragment retrieval system

The input to the proposed retrieval system in Fig. 3 is a set of collected evidence E. This model is composed of three main components: search, matching and aggregation. They interact with each other in the following steps:

Initially, the collected evidence goes through the system one piece at a time. Given the knowledge base, the search component efficiently identifies the position of the required concept/predicate in the corresponding taxonomy. After that, the resulting information (part of the taxonomy) and the piece of evidence are both fed into the matching component. Atomic matching degrees are generated by this process and are then aggregated together, deriving an overall degree of match for the given evidence. Next, the degree of match is fed back to the search component to guide further search. During this process, the search mechanism keeps those concepts/predicaters whose degree of match is above a predefined threshold for further search only. At the next search stage, it focuses on picking up model fragments which are related to the selected concepts/predicaters. These retrieved model fragments are then stored in a candidate pool for final creation of the possible scenarios along with other candidates to be retrieved by repeating this cycle.

The above process continues until all evidence in E is examined sequentially. The aggregation component will then be employed to update and calculate the final RSV of each candidate model fragment. Finally, the output of this system is an ordered set S of retrieved model fragments, which have been at least partially instantiated in relation to the collected evidence. Technical details for implementing the main steps are described in the following subsections.

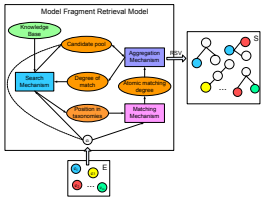

Fig. 3. Model fragment retrieval system

D. Searching through taxonomies

In the IR literature, most efforts have been devoted to calculating the degree of relevance with limited consideration of the search process for efficient matching. They often store all the documents in the disk; then, given a query, check every document in the disk one by one. This exhaustive process works, but it is inefficient and can be computationally prohibitive. However, fetching the "right" information from storage is not an easy task. This is because the knowledge base may well contain several hundreds or sometimes even tens of thousands of model fragments along with further basic concepts/predicates that describe states and events appearing in the model fragments that are structured in weighted taxonomies. The efficiency of fuzzy model fragment retrieval is therefore crucially related to the systems's capability to search for the most relevant model fragments on the basis of RSV. For this purpose, a search algorithm is developed to improve the efficiency of the overall system.

Algorithm : SearchTaxonomy(E, KB)

1) Extract the atomic concepts in E
2) Count the number of occurrences of each atomic concept
3) Store the atomic concepts (c) in a queue Q, in decreasing order of the numbers of their occurrences
4) Take the first node c_i of Q
   - Assign by expert which taxonomies (T_i) c_i might belong to
   - Perform the *breadth-first-search*(c_i, T_i) and return its position in T_i or "Not found"
5) Repeat Step 4 until Q is empty

For example, given a set of evidence $E = \{e_1, e_2, e_3\}$ where

e_1 = (height(human_being(John)) = tall) $\longrightarrow$ somewhat certain
e_2 = (height(human_being(Dave)) = short) $\longrightarrow$ good chance
e_3 = (employ(human_being(John), human_being(Dave)) = true) $\longrightarrow$ extremely likely

It is obvious that the atomic concepts in E are *human_being*, *height* and *employ* and the number of their occurrences are 4, 2 and 1 respectively. Thus, a queue Q is constructed as:

$$Q = \{human\_being, height, employ\}$$

The predefined weighted taxonomies are employed to guide the retrieval system to more relevant part of the knowledge base. Take the first concept in Q, *human_being*, expertise is required to identify which taxonomy it might belong to. The search starts from the root node of a selected taxonomy and breadth-first search is employed to identify the position of the required concept in the given taxonomy.

In order to reduce the number of search iterations and improve the retrieval efficiency, when retrieving a concept which has already existed in E, rather than searching it from scratch each time, it has a natural appeal to save the previous retrieval results in the cache, such that, the search results of these concepts can be retrieved directly for the next iteration. This is why the number of occurrences is counted in the algorithm. More broadly, the results of most frequently used concepts/predicates such as *human_being*, *place*, *time* can be recorded in one mapping table and stored in the knowledge base. Also, these high-frequency retrieval concepts are ordered according to their retrieval frequency. The more frequently they are required, the higher ranking they appear in the mapping table, such that retrieval time is reduced. Therefore, repeated search and retrieval for those most frequent concepts can be avoided.

Note that, the above search mechanism is implemented based on exact word matching, it suffers from the problem that concepts which are semantically similar might be missed. In order to overcome this problem, an enhanced semantic matching mechanism will be integrated in next step.

E. Semantic matching

In this step, the concepts/predicates appearing in the collected atomic evidence are matched with predefined weighted taxonomies at a semantic level in order to calculate the semantic similarity S_v between them.

Recently, many approaches concerning semantic matching have been proposed, represented by the work done is semantic web area. For example, Giunchiglia [10] presented a structure level semantic matching algorithm to find semantic correspondences between elements of two graph-like structures. Another group of approaches focus on the extension of the original query terms. For instance, Miyamoto [15] proposed a fuzzy association mechanism for generating a fuzzy thesaurus for all pairs of keywords in a given set of documents based on their frequency of occurrence. More recently, some approaches start to move away from word occurrences but onto concept occurrences [9].

In this paper, given a weighted taxonomy T, so long as two concepts have overlapping semantics, the semantic matching degree is set to be greater than 0. For example, in Fig. 1, the "Coke" and "Liquid" should have a non-zero semantic matching degree, since they share something in common in respect of their ingredients. The semantic matching degree between any two nodes within a given taxonomy T can be derived via the weighted length of the shortest path connecting them. This is because the arc weight indicates the semantic distance between two connecting nodes. That is, the semantic similarity between c and c' is defined by:

$$S_v(c, c') = \left(1 - \frac{N_p}{N_t}\right) * \frac{S}{N_p} * G^{|d_c - d_{c'}|} \quad (2)$$

where N_p is the number of edges of the shortest path connecting c and c' in T and N_t is the number of total edges of T. Obviously, $(1 - \frac{N_p}{N_t})$ indicates the semantic distance between c and c'. S represents the sum of arc weights attached on the shortest path. $\frac{S}{N_p}$ stands for the normalized weight of the shortest path. Also, the level difference is taken into consideration as $\mid d_c - d_{c'} \mid$ is the absolute level difference between c and c' in T. G is a parameter which inversely signifies the importance of level difference in the

sematic matching. The larger its value the less important the level difference.

This definition captures intuition well. Take Fig. 1 for example, suppose that $G = 0.95$, then the semantic matching degree between "Coke" and "Liquid" is obtained by using Equation (2):

$$S_v(\text{Coke}, \text{Liquid}) = \left(1 - \frac{2}{11}\right) * \frac{0.95 + 1.0}{2} * 0.95^2 = 0.72$$

Through the use of this similarity measure, once the required concept, c, is identified in T, its semantic similarity with other concepts in T can be obtained. Those concepts that have a similarity degree larger than a given threshold will be selected. It is inefficient to traverse the whole taxonomy, since if the similarity between c and a node in T is smaller than the threshold and this node is located at the same level as c or lower, then the whole branch can be pruned.

F. Finding relevant model fragments

Based on the above semantic match, an involved concept c in a given piece of evidence can be expanded to a broader set, covering those concepts that all have a semantic similarity degree with c that is larger than the threshold. As a piece of evidence may involve nested proposition (or instantiated predicates), each atomic concept/predicate should be expanded individually. After that, with the help of the predefined hash table, the index numbers of those model fragments which are relevant to the interested concepts/predicates can be easily determined. In addition, the semantic similarity degree can also be attached to each of such model fragments.

This step is set to retrieve model fragments which are relevant to the whole evidence proposition rather than atomic elements. Thus, intersection ($\cap$) is imposed over those individually determined model fragment sets to retrieve the model fragments that jointly involve all concepts/predicates in the given piece of evidence.

Of course, the semantic similarity degrees associated with the atomic predicates should also be aggregated. The overall semantic similarity degree between the retrieved model fragment (ϕ') and evidence (ϕ) can be obtained by aggregating the semantic similarity values of atomic predicates.

$$S_v(\phi, \phi') = S_v(p_1, p_1') \oplus \cdots \oplus S_v(p_m, p_m') \tag{3}$$

Here, the aggregation operator ($\oplus$) is interpreted as *average* to avoid the semantic similarity degree rapidly degrading at the very beginning, otherwise very limited candidate model fragments can be forwarded to the following steps.

In summary, given a piece of evidence e_i, at the end of this step, those model fragments that possess an overall proposition semantic similarity with e_i which is larger than a predefined threshold are selected and stored in a candidate pool for further reduction.

G. Fuzzy set matching

According to the representation of fuzzy model fragments in section II-A.2, the distribution component consists of a set of rules of the following form:

$$\begin{array}{l}\text{If } \phi_1^s \text{ is } v_{11}^s \text{ and } \cdots \text{ and } \phi_n^s \text{ is } v_{1n}^s \text{ Then } \phi_1^t \text{ is } v_{11}^t \text{ and } \cdots \text{ and } \phi_m^t \text{ is } v_{1m}^t \\ \cdots\cdots \\ \text{If } \phi_1^s \text{ is } v_{r1}^s \text{ and } \cdots \text{ and } \phi_n^s \text{ is } v_{rn}^s \text{ Then } \phi_1^t \text{ is } v_{r1}^t \text{ and } \cdots \text{ and } \phi_m^t \text{ is } v_{rm}^t\end{array}$$

Without losing generality, each rule has n antecedent predicates and m consequent predicates. $\phi_1^s, \cdots, \phi_n^s$ and $\phi_1^t, \cdots, \phi_m^t$ are predicates/functions and $v_{11}^s, \cdots, v_{rn}^s$ and $v_{11}^t, \cdots, v_{rm}^t$ are their established values. The predicates in fuzzy model fragments are allowed to be precise or imprecise or mixed. The precise predicate can only be evaluated to be true or false, whereas the value of an imprecise predicate/function may be represented by a fuzzy set defined on a suitable universe of discourse.

Given an atomic evidence, ϕ is v, the semantic match between ϕ and ϕ^s/ϕ^t is performed in the last step. In this step, the similarity of predicate values is calculated. If ϕ is a precise predicate, the value of ϕ always exactly matches the value of ϕ^s/ϕ^t, the matching degree between them is either 1 or 0. For an imprecise predicate or function, the fuzzy set matching degree takes values from a continuous range $[0, 1]$, where 1 indicates two predicates are completely matched and 0 indicates they are not matched at all. Note that, according to the definition of fuzzy taxonomy in [8], although the fuzzy concepts involved may not be exactly the same, their values which are represented by fuzzy sets defined in the same normalized universe of discourse are still comparable. The degree of overlap between a pair of such fuzzy sets reflects the similarity between them. Thus, the fuzzy set matching degree, S_f , between the evidence and a model fragment can be obtained.

In this paper, the technique reported in [14] is used to measure fuzzy set similarity. This measure of similarity is based upon two measures, possibility Pos and necessity Nec. This method can be used consistently for all possible combinations of precise and imprecise propositions, regarding pieces of evidence and model fragments. Formally, the similarity measure, S_f, is defined by:

$$Pos(v_m|v_e) = \max(\min(\mu_{v_m}(u), \mu_{v_e}(u))) \quad \forall u \in U \tag{4}$$

$$Nec(v_m|v_e) = 1 - \max(\min(1 - \mu_{v_m}(u), \mu_{v_e}(u))) \ \forall u \in U \tag{5}$$

$$S_f = \begin{cases} Pos(v_m \mid v_e) & \text{if } Nec(v_m \mid v_e) > 0.5 \\ (Nec(v_m \mid v_e) + 0.5) \times Pos(v_m \mid v_e) & \text{else} \end{cases} \tag{6}$$

where U is the universe of discourse and Pos the maximum value among the intersection point of v_e and v_m (representing to what extent v_e and v_m overlap).

Based on the above semantic and fuzzy set matching, the overall fuzzy matching degree between "ϕ is v" and "ϕ' is v'" are then aggregated by using the *product* operator:

$$S(\phi : v, \phi' : v') = S_v(\phi, \phi') \oplus S_f(\phi : v, \phi' : v') \tag{7}$$

Here, the *product* operator is adapted for aggregation because crisp predicates are often involved, such that the S_f of crisp predicates is either 1 or 0. It is obvious that if $S_f = 1$, then the use of other aggregation operators such as *plus* or *Max* may lead to the S_f of crisp predicates dominating the overall aggregation process and weakening the effect of sematic similarity S_v. On the other hand, if $S_f = 0$, the

overall fuzzy matching degree tends to be 0 by using *product* operator. This clearly reflects the intuition well.

H. Rule level aggregation

Since each model fragment employs a set of generic rules, given an atomic piece of evidence e, more than one rule can usually be fired with different fuzzy matching degrees. This step aims to combine the individual fuzzy matching degrees to derive an overall relevant degree for a model fragment. In other words, after this step, for a given e, each model fragment in the candidate pool will have an overall relevant degree attached, which signifies how relevant the model fragment is to the given evidence.

The individual fuzzy match degree of the fired rules in one model fragment, $S_1, S_2, \cdots, S_n$, is herein aggregated as:

$$f_i = f_{i-1} + S_i - f_{i-1} \times S_i \tag{8}$$

where $f_0 = S_0 = 0$ and $i = 1, 2, \ldots, n$, with n being the number of fired rules in one model fragment and f_n being the overall relevant degree of a candidate model fragment. The resulting value is bounded by the maximum individual fuzzy match degree and 1. Again, this well reflects the intuition in that the more rules are fired with respect to a single model fragment, the higher relevant degree exists between this fragment and the given piece of evidence.

I. Model fragment level aggregation

The above steps will be iteratively carried out until all atomic pieces of evidence in E have been examined. For each e_i, a set of relevant model fragments are obtained and each model fragment will attach a relevant degree (the output of last step) with e_i. However, a model fragment normally consists of several atomic predicates which are connected by conjunctive operators, whilst it has been assumed that each atomic piece of evidence can only activate one atomic predicate/function at a time. In addition, each atomic evidence has a confidence weight associated, denoting the degree of certainty of that piece of evidence. Importantly, the confidence weights are generally given in linguistic terms for the application problem at hand. Therefore, if a single model fragment is matched by more than one atomic evidence, their individual relevant degrees should be aggregated.

For computational simplicity, in this work, a finite and totally ordered linguistic term set, $L = \{l_i\}$, $i \in H = \{0, \cdots, T\}$, is employed to cover permissible linguistic terms. That is, it takes the *ordinal fuzzy linguistic approach* [4], [6], where any linguistic term l_i in L represents a possible value for a linguistic variable. As with the existing work, the term set L must satisfy [4]:

- Totally ordered: $l_i \geq l_j$ if $i \geq j$.
- Symmetry: $Neg(l_i) = l_j$ if $j = T - i$.
- Maximization: $Max(l_i, l_j) = l_i$ if $l_i \geq l_j$.
- Minimization: $Min(l_i, l_j) = l_i$ if $l_i \leq l_j$.

In particular, the confidence weights are described using a set of linguistic probability terms L, as defined as follows:

$$L = \{l_0 = Impossible, l_1 = Extremely\_unlikely, l_2 = Slim\_chance, l_3 = Small\_chance, l_4 = Somewhat\_certain, l_5 = Good\_chance, l_6 = Most\_likely, l_7 = Extremely\_likely, l_8 = Certain\}$$

This final step is desired to combine the numerical relevant degree and linguistic confidence weight to create the final RSV for a candidate model fragment:

$$RSV_{MF_j} = f(\langle l_1, R_{MF_{j_1}} \rangle, \langle l_2, R_{MF_{j_2}} \rangle, \cdots, \langle l_n, R_{MF_{j_n}} \rangle)$$

where j is the index number of a model fragment in the candidate pool, n is the total number of atomic pieces of evidence in E, l_i is the confidence weight of the i^{th} piece of evidence in E and $R_{MF_{j_n}}$ stands for the relevant degree of j^{th} model fragment with respect to the n^{th} atomic evidence.

Traditionally, when a given pattern expresses the conjunction of certain elementary requirements, ***Min*** operation is typically performed in aggregation. However, the use of ***Min*** often leads to a very small value dominating the aggregation function and forcing a conclusion to be drawn based on the least relevant proposition, which is just the opposite of what is desired by the user for the present application.

Instead, the calculation of the final RSV for a candidate model fragment is herein done by the following procedure. First, calculate the individual relevant degree of MF_i with e_i using the above steps. Note that, if the j^{th} model fragment has not been matched by the i^{th} evidence, $R_{MF_{j_i}} = 0$. Second, the individual relevant degrees are aggregated using the Induced OWA (IOWA) Operators [19] such as:

$$RSV_{MF_j} = f_{\boldsymbol{\omega}}(\langle l_1, R_{MF_{j_1}} \rangle, \ldots, \langle l_n, R_{MF_{j_n}} \rangle) = \boldsymbol{\omega}^T B \tag{9}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_n)$ is the weighting vector with $\omega_i \in [0, 1]$ and $\sum_{i=1}^{n} \omega_i = 1$. This is well suited to the present work, since the relevant degrees to be aggregated can be ordered according to their attached linguistic confidence weights. Indeed, the confidence weight l_i of this research is the so-called order-inducing variable in [19], which is used to reorder the aggregated objects, $R_{MF_{j_i}}$ is the argument variable, and B is a reordered argument vector so that b_k is the $R_{MF_{j_i}}$ value of the aggregated object having the k^{th} largest l value. Third, construct the weighting vector $\boldsymbol{\omega}$. Normally, the weighted vector is assigned by expert or by learning from historical records. In this research, the weights in $\boldsymbol{\omega}$ reflect the confidence degrees; it is therefore nature to normalize the index number of each evidence's confidence weight in L to construct $\boldsymbol{\omega}$.

This aggregation scheme once again reflects the intuition that the more evidence there exists to support a model fragment, the higher is the relevance of that model fragment likely to involve in the scenario to build. The final RSV is used as a criterion to select those most relevant model fragments in the candidate pool. This hierarchical aggregation procedure is summarized in Fig. 4. For this to work, a threshold is predefined to indicate the minimum acceptance level for the RSV of a model fragment, to be selected. Those selected model fragments are then used to construct the scenario description.

IV. Illustrative Example

In the wake of recent terrorist atrocities, intelligence experts have commented that failures in detecting terrorist activities are not so much due to a lack of data, as they are due to difficulties in relating and interpreting the available

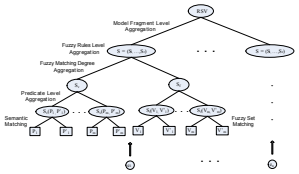

Fig. 4. Hierarchical aggregation structure

intelligence data. For example, it is obvious that many explosive ingredients and liquids can be combined to create homemade liquid bomb. However, there are a lot of explosive chemicals that can be concocted from some very common items such as perfumes, drain cleaner and batteries and they are innocent in themselves. This makes it very difficult for intelligence analysts to detect a plausible threat. The proposed model fragment retrieval system is designed to support the selection of a broader set of relevant model fragments, by matching observations with a knowledge base when both imprecise and uncertain information is obtained, in order to depict the likely scenarios which may well link such instantiated fragments together. The retrieval process described in the preceding section is illustrated by addressing a simply application problem.

Assume that a suspect was trying to bring a bottle of coke boarding a flight and a small bag of hair dyes was also found in his suitcase. These two pieces of evidence were observed by police and stored in $E = \{e_1, e_2\}$:

$e_1 = \phi_1^e$: Amount(Coke(a)) = many $\rightarrow$ most likely (l_6)
$e_2 = \phi_2^e$: Amount(Hair_dyes(b)) = a_few $\rightarrow$ somewhat certain (l_4)

Further, supposed that the knowledge base contains the following two (candidate) model fragments MF_1 and MF_2:

If $\{\phi_{11}^s$: `Amount(Liquid(X))`, ϕ_{21}^s : `Amount(Hydrogen_peroxide(Y))`$\}$
Assuming $\{\phi_{31}^s$: `Mix(Liquid(X), Hydrogen_peroxide(Y))`$\}$
Then $\{\phi_{11}^t$: `Liquid_bmob(B)`$\}$
Distribution `Liquid_bmob(B)` {
`Rule 1: a_lot, a_few, true →` **true: extremely_likely** (l_7)
`Rule 2: a_few, a_lot, true →` **true: slim_chance** (l_2)
`Rule 3: a_number_of, a_number_of, true →`
true: somewhat_certian (l_4)$\}$ (MF_1)

If $\{\phi_{12}^s$: `Pregnant_woman(X)`$\}$
Assuming $\{\phi_{22}^s$: `Amount(intake(caffeine(Y)))`$\}$
Then $\{\phi_{12}^t$: `Cause_miscarriage(X)`$\}$
Distribution `Cause_miscarriage(X)` {
`Rule 1: true, a_lot →` **true: extremely_likely** (l_7)
`Rule 2: true, a_number_of →` **true: small_chance** (l_3)
`Rule 3: true, a_few →` **true: slim_chance** (l_2)$\}$ (MF_2)

Given such two candidate model fragments and a predefined fuzzy taxonomy (as shown in Fig. 1), the following model fragment retrieval process then follows:

To start, e_1 is firstly fed into the model fragment retrieval system, the involved concept *Coke* is suggested to search from the *Substance* taxonomy (see Fig. 1). This leads to the following computation:

1) Calculate the semantic similarities: Suppose that G is set to be 0.95 and the threshold to 0.7, the results of applying Equation (2) are listed in the second column of Table I which

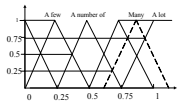

Fig. 5. Fuzzy set matching

lists the obtained semantic similarities between the desired concepts in the set of collected evidence and other concepts in the given taxonomy. After this, the original concept *Coke* in e_1 can be expanded to be as:

$$Exp(Coke) = \{Liquid, Drink, Caffeine, Coke\}$$

TABLE I
RESULTS OF SEMANTIC MATCH

	Coke	Hair dyes
Substance	0.65	0.49
Explosive integrate	0.52	0.61
Liquid	0.72	0.48
Hydrogen peroxide	0.46	0.73
Sulphuric acid	0.46	0.56
Oil	0.52	0.38
Drink	0.82	0.45
Hair dyes	0.40	1
Coke	1	0.40
Tea	0.68	0.38
Caffeine	0.73	—
Sugar	0.69	—

Applying the forward chaining inference mechanism leads to both model fragments MF_1 and MF_2 being selected. For illustrative simplicity, assume that the expanded sets for other concepts *Amount* and *Intake* return themselves, i.e. $S_v(Amount, Amount) = 1$ and $S_v(Intake, Intake) = 1$. From this, the overall semantic similarities of these selected model fragments with e_1 can be obtained:

$$\begin{aligned} S_v(\phi_{11}^s, \phi_1^e) &= (S_v(Amount, Amount) + S_v(Coke, Liquid))/2 = 0.86 \\ S_v(\phi_{22}^s, \phi_1^e) &= (S_v(Amount, Amount) + S_v(Intake, Intake) + \\ &\quad S_v(Coke, Caffaine))/3 = 0.91 \end{aligned}$$

2) Calculate the similarity of predicate/function values: Suppose that the fuzzy variable *Amount* is defined as per Fig. 5. Only two rules (Rule 1 and Rule 3) in MF_1 and two rules (Rule 1 and Rule 2) in MF_2 are fired by e_1 respectively. Applying Equation(6) to the above leads to:

$$\begin{aligned} &S_f(\phi_{11}^s : \text{a\_lot},\ \phi_1^e : \text{many}) = 0.5625 \\ &S_f(\phi_{11}^s : \text{a\_number\_of},\ \phi_1^e : \text{many}) = 0.125 \\ &S_f(\phi_{22}^s : \text{a\_lot},\ \phi_1^e : \text{many}) = 0.5625 \\ &S_f(\phi_{22}^s : \text{a\_number\_of},\ \phi_1^e : \text{many}) = 0.125 \end{aligned}$$

3) Calculate the overall fuzzy matching degree between e_1 and MF_1: This is done by using Equation (7):

$$\begin{aligned} &S(\phi_{11}^s : \text{a\_lot},\ \phi_1^e : \text{many}) \\ &= S_v(\phi_{11}^s, \phi_1^e) \times S_f(\phi_{11}^s : \text{a\_lot}, \phi_1^e : \text{many}) = 0.48 \\ &S(\phi_{11}^s : \text{a\_number\_of},\ \phi_1^e : \text{many}) \\ &= S_v(\phi_{11}^s, \phi_1^e) \times S_f(\phi_{11}^s : \text{a\_number\_of}, \phi_1^e : \text{many}) = 0.11 \end{aligned}$$

4) Calculate the overall relevant degree between MF_1 and e_1: This is computed by using Equation (8):

$$\begin{aligned} R_{MF_{1_1}} &= S(\phi_{11}^s : \text{a\_lot},\ \phi_1^e : \text{many}) + \\ &\quad S(\phi_{11}^s : \text{a\_number\_of},\ \phi_1^e : \text{many}) - \\ &\quad S(\phi_{11}^s : \text{a\_lot},\ \phi_1^e : \text{many}) \times \\ &\quad S(\phi_{11}^s : \text{a\_number\_of},\ \phi_1^e : \text{many}) = 0.54 \end{aligned}$$

Similarly, repeat steps 2 to 4, the overall relevant degree between MF_2 and e_1 results:

$$\begin{aligned} R_{MF_{2_1}} &= S(\phi_{22}^s : \text{a\_lot},\ \phi_1^e : \text{many}) + \\ &\quad S(\phi_{22}^s : \text{a\_number\_of},\ \phi_1^e : \text{many}) - \\ &\quad S(\phi_{22}^s : \text{a\_lot},\ \phi_1^e : \text{many}) \times \\ &\quad S(\phi_{22}^s : \text{a\_number\_of},\ \phi_1^e : \text{many}) = 0.56 \end{aligned}$$

5) Repeat the above steps until all evidence in E are examined: As given in Table I (the third column), the expanded set for *Hair dyes* in e_2 is:

Exp(*Hair dyes*) = {*Hair dyes*, *Hydrogen peroxide*}

This is because the sematic similarities of between these two concepts and *Hair dyes* are larger than 0.7. Note that, since the concepts *Coke* and *Hair dyes* are both at the same level and $S_v(Coke, Hairdyes) < 0.7$, the child node of *Coke* has no chance to get a similarity value S_v with *Hair dyes* which is larger than 0.7, the whole branch of *Coke* can therefore be removed. Only one model fragment (MF_1) is thus activated by e_2. From this, the overall relevant degree between MF_1 and e_2 can be obtained:

$$\begin{aligned} R_{MF_{1_2}} &= S(\phi^s_{21}: \text{a\_few},\ \phi^e_2: \text{a\_few}) + \\ &\quad S(\phi^s_{21}: \text{a\_number\_of},\ \phi^e_2: \text{a\_few}) - \\ &\quad S(\phi^s_{21}: \text{a\_few},\ \phi^e_2: \text{a\_few}) \times \\ &\quad S(\phi^s_{21}: \text{a\_number\_of},\ \phi^e_2: \text{a\_few}) = 0.93 \end{aligned}$$

6) Calculate the final RSV for each candidate model fragment: This is accomplished by using Equation (9). In this example, MF_1 is fired by both e_1 and e_2 with relevant degrees $R_{MF_{1_1}} = 0.54$ and $R_{MF_{1_2}} = 0.93$, respectively, and MF_2 is only fired by e_1 with relevant degree $R_{MF_{2_1}} = 0.56$. It is known that the confidence weight of e_1 is most_likely (l_6) and that of e_2 is somewhat_certian (l_4). Thus, the final RSVs of MF_1 and MF_2 can be derived as follows:

The first step is to order the aggregation pairs, $\langle l_i, R_{MF_{j_i}}\rangle$, based upon the ordering inducing variable l_i. The reordered aggregation pairs for MF_1 are $\langle l_6, 0.54\rangle, \langle l_4, 0.93\rangle$ and this order leads to the ordered argument vector $B = [0.54\ 0.93]$. In addition, the weighting vector $\boldsymbol{\omega}$ is constructed as $\omega_1 = 6/6{+}4 = 0.6$ and $\omega_2 = 4/6{+}4 = 0.4$. Thus, $\boldsymbol{\omega}^T = [0.6\ 0.4]$ and the final RSVs of MF_1 and MF_2 are obtained as:

$$\begin{aligned} RSV_{MF_1} &= f_{\boldsymbol{\omega}}(\langle l_6, R_{MF_{1_1}}\rangle, \langle l_4, R_{MF_{1_2}}\rangle) = 0.69 \\ RSV_{MF_2} &= f_{\boldsymbol{\omega}}(\langle l_6, R_{MF_{2_1}}\rangle, \langle l_4, R_{MF_{2_2}}\rangle) = 0.34 \end{aligned}$$

Obviously, given the collected evidence E, boolean retrieval model will ignore both model fragments MF_1 and MF_2. However, if the suspect mixes the hair dyes and coke together in a appropriate proportion, a liquid bomb might be produced. Boolean retrieval model is incapable of detecting and constructing this plausible threat, whilst the present approach can assist in making such difficult decisions.

V. Conclusions

This paper has proposed an approach to develop flexible model fragment retrieval systems which are capable of performing partial matches among pieces of imprecise and uncertain information. For a given set of collected evidence, the retrieval procedure carries out following distinctive tasks: 1) semantic matching of atomic evidence predicates. 2) fuzzy set matching of predicate values. 3) aggregation of individual fuzzy matching degrees of those fired fuzzy rules in each candidate model fragment. 4) aggregation of the final RSV of each model fragment if it has been fired by more than one atomic piece of evidence. Compared with boolean retrieval, if a match between the collected evidence and the knowledge base cannot be established precisely with full certainty, then no result may be produced. However, the approach described herein can pick up those relevant model fragments whose RSV exceeds a given threshold.

In future work, an inference mechanism will be developed to synthesise and store all combinations of instances of these selected model fragments. For the application problems considered, this will allow scenarios depicting plausible terrorist attacks that might never be considered by analysts to be created. Another interesting piece of further research is to devised a mechanism to propagate computed fuzzy matching degrees and linguistic probabilities from individual model fragments to their related ones. How to best decide the threshold value required by the retrieved process for a fresh application domain also remains an important piece of further research. Also, the numeric relevant weights in the present weighted taxonomy may be more naturally represented by fuzzy sets, especially in the present application domain.

References

[1] G. Bordogna and G. Pasi. A fuzzy linguistic approach generalizing boolean information retrieval: A model and its evaluation. *The American Society for Information Science*, 44(2):70–82, 1993.

[2] G. Bordogna and G. Pasi. Linguistic aggregation operators of selection criteria in fuzzy information retrieval. *International journal of intelligent systems*, 10(32):233–248, 1995.

[3] G. Bordogna and G. Pasi. Modeling vagueness in information retrieval. In *Proceedings of the Third European Summer-School on Lectures on Information Retrieval-Revised Lectures*, pages 207–241, 2000.

[4] M. Delgado, F.Herrera, E. Herrera-Viedma, and L. Martínez. Combining numerical and linguistic information in group decision making. *Information Science*, 107(1):177–194, 1998.

[5] D. Dubois, H. Prade, and C. Testemale. Weighted fuzzy pattern matching. *Fuzzy Sets and Systems*, 28:313–331, 1988.

[6] E.Herrera-Viedma. Modeling the retrieval process of an information retrieval system using an ordinal fuzzy linguistic approach. *Journal of the American Society of Information Science*, 52(6):460–475, 2001.

[7] B. Falkenhainer and K. Forbus. Compositional modelling: Finding the right model for the job. *Artificial Intelligence*, 51:95–143, 1991.

[8] X. Fu, Q. Shen, and R. Zhao. Towards fuzzy compositional modelling. In *Proceedings of the 16th International Conference on Fuzzy Systems*, pages 1233–1238, 2007.

[9] P. J. Garcés, J. A. Olivas, and F. P. Romero. Concept-matching ir systems versus word-matching information retrieval systems: Considering fuzzy interrelations for indexing web pages. *Journal of the American Society for Information Science and Technology*, 57(4):564–576, 2006.

[10] F. Giunchiglia, M. Yatskevich, and E. Giunchiglia. Efficient semantic matching. In *The Semantic Web: Research and Application*, pages 272–289. Springer Berlin / Heidelberg.

[11] J. Halliwell and Q. Shen. Linguistic probabilities: theory and application. *To appear in Soft Computing*, 2007.

[12] J. Keppens and Q. Shen. On compositional modelling. *Knowledge Engineering Reivew*, 16(2):157–200, 2001.

[13] D. Kraft and D. Buell. Fuzzy sets and generalized boolean retrieval systems. *International Journal of Man-Machine Studies*, 19(1):45–56, 1983.

[14] K. Leung and W. Lam. Fuzzy concepts in expert systems. *IEEE Computer*, 21(8):43–56, 1988.

[15] S. Miyamoto. Information retrieval based on fuzzy associations. *Fuzzy Sets Syst.*, 38(2):191–205, 1990.

[16] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[17] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, 1988.

[18] R. R. Yager. Second order structures in multi-criteria decision making. *Int. J. Man-Mach. Stud.*, 36(4):553–570, 1992.

[19] R. R. Yager and D. P. Filev. Induced ordered weighted averaging operators. *IEEE Transaction on Systems, Man and Cybernetics*, 29:141–150, 1999.