Augmenting the Distributed Evaluation of Path Queries with Information Granules

Davide Bacciu[1], Alessio Botta[1], and Dan Stefanescu[2]

[1] IMT Lucca Institute for Advance Studies.
Piazza San Ponziano 6, 55100 Lucca, Italy
d.bacciu,a.botta@imtlucca.it
[2] Mathematics and Computer Science Department, Suffolk University.
Beacon Hill, Boston, MA, USA
dan@mcs.suffolk.edu

1 Introduction

The advent of ubiquitous fast networks, cheap storage and processing cycles allows for the accumulation, organization and access of large collections of data. In many areas such as communications and traffic networks, biological data management, cartography and web information systems large databases are represented as labeled graphs for which regular path queries (RPQs) [1] represent a common and convenient way to access and extract knowledge. Users can further specify the required knowledge by expressing their queries in terms of weighted RPQs (essentially weighted automata) and requesting the return of the cheapest such specified paths. Recent work [2, 3] explored computational aspects of evaluating regular path queries on large, weighted distributed data-graphs and, in particular, single and multiple source distributed evaluation, termination detection, fault tolerance and computation in grid environments. In this work we further look to enhance the practicality of extracting information from data-graphs by augmenting the expressive power with the use of *information granules* [4].

For instance, consider the example of a *spatial network*, e.g., a road map [3]. A typical RPQ for this application domain consists in defining the initial location, the destination, and some intermediate locations along the desired paths. Further, symbols can be associated with edges to define a specific kind of connection (road) between locations (cities). An RPQ provides us with enough descriptive power to express a request such as "*I would like to go from Pisa to Florence via Lucca, following a road between Pisa and Lucca, and following a road or a highway between Lucca and Florence*". One can further limit the paths returned by the evaluation of the aforementioned query by designing a weighted RPQ, but this approach is too rigid, lacks convenience and will not be able to allow requests such as "*I would like to go from Pisa to Florence via Lucca, following a road between Pisa and Lucca in at most 20 minutes, and following a highway in about 30 minutes or a road in at least 40 minutes between Lucca and Florence*".

To allow for such requests we propose to enhance RPQs with elements of *granular computing* [4], a theory that deals with operations performed over information granules, rather than singular and exact values. Granular computing models the abstraction process of the human mind, and allows to associate a *semantic meaning* to data, i.e., to "*compute with words*" [5]. In the literature, many theories to represent information granules have been developed. Fuzzy set theory (FST) [5] is the most established theory of information granulation. The basic computational units of FST are *fuzzy sets*, i.e. sets with elements whose *degree of membership* ranges in the $[0, 1]$ real interval. For instance, the imprecise time distances expressed by "*at most 20 minutes*", "*about 30 minutes*" and "*at least 40 minutes*" of the previous example can be easily modeled by the fuzzy sets depicted in Fig. 1. In our work, we represent information granules in the form of fuzzy sets and, more specifically, *fuzzy quantities*, i.e. fuzzy sets defined over a real-valued universe of discourse [6, 7].

2 The Model

2.1 The Basic Model

Let us consider a data-graph DB that can be represented as a weighted and labeled graph $DB = (V, E, \mu, \omega)$, where $V = \{o_0, \ldots, o_N\}$ is the set of vertices representing data-graph objects and $E \subseteq V \times V$ is the set of edges. The functions $\mu : E \rightarrow \Delta$ and $\omega : E \rightarrow \mathbb{K}$ assign the edge labels and weights, respectively. The signature Δ defines an alphabet of symbols in a specific domain. For instance, in the domain of spatial networks, values in Δ can be *road to* C_i, *highway to* C_i, *freeway to* C_i, *bridge to* C_i, etc. The weight set $\mathbb{K}$ contains elements in the domain knowledge associated to the edges. More formally, a graph edge $e_j = (o_j, o'_j)$ represents a relationship between objects o_j and o'_j, identified by $\mu(e_j) \in \Delta$ and tied to the domain knowledge associated to the information granule $\omega(e_j) \in \mathbb{K}$. Figure 2(a) shows a sample spatial network

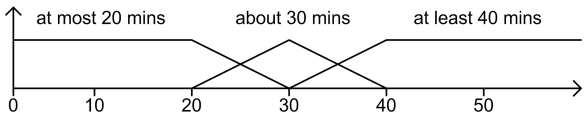

Fig. 1. Fuzzy quantities modeling imprecise time distances.

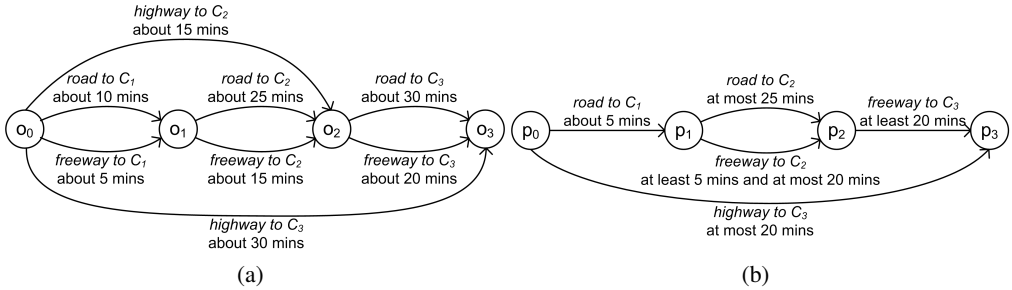

Fig. 2. Sample (a) data-graph and (b) RPQ with fuzzy weights.

weighted by fuzzy quantities that model imprecise time distances. The weights defined by ω over $\mathbb{K}$ are used to compute a cost $c(\pi)$ of each path π starting from vertex o_0 and ending at vertex o_F. The cost $c(\pi)$ is computed in a general domain $\mathbb{C}$, whose nature depends on the specific application. In order to compute $c(\pi)$, we need means for aggregating the costs of each edge. We call this operator the *cost sum*, denoted as $\oplus_{\mathbb{C}} : \mathbb{C} \times \mathbb{C} \rightarrow \mathbb{C}$. Moreover, we need an order relation $\preceq_{\mathbb{C}}$ to establish an ordering between the costs in $\mathbb{C}$, so to be able to determine cheap paths. For instance, in the simplest cases [2, 3], both weights and costs are real numbers, modeling, e.g., the length in miles of a road and the total length of the trip. Formally, we have $\mathbb{K} \equiv \mathbb{C} \equiv \mathbb{R}$, the cost-sum $\oplus_{\mathbb{C}}$ is the real $+_{\mathbb{R}}$ operator, and the $\preceq_{\mathbb{C}}$ is the standard ordering relation $\leq_{\mathbb{R}}$ between reals.

An RPQ on such a DB is described by a finite state automaton (FSA) $A = (P, \Delta, \tau, p_0, P_F)$, where P is the set of states, Δ is the signature, τ is the transition relation, p_0 is the initial state, and P_F is the set of final states. Each edge in the automaton identifies a query term and every path leading from the initial state p_0 to a final state $p_f \in P_F$ determines an admissible instance of the query, i.e., an acceptable path.

The original algorithm looks for the cheapest acceptable paths defined by an RPQ A over a DB distributed on a grid of machines [3]. The query matching procedure consists in finding sub-graphs of the DB that satisfy the RPQ. Roughly speaking, a matching sub-graph is such if its edge labels match with the transition labels of any of the acceptable paths defined in the RPQ. The algorithm starts from a root vertex o_0 and proceeds by incrementally building the set of optimal acceptable solutions in a distributed fashion. Each time a new (partially) acceptable path is found, its (partial) cost is computed by aggregating the costs of the edges via $\oplus_{\mathbb{C}}$. If the newly found path is better than the (possibly) already existing one, then the former replaces the latter in the set of optimal acceptable solutions. Evaluation of the cheapest path is performed via $\preceq_{\mathbb{C}}$. At each time instant, each machine of the grid is aware of the best partial solutions discovered up to that instant which contain at least a vertex stored in its local memory. Although the algorithm is based on a greedy strategy, it is proved to return the optimal complete paths. Further, the algorithm includes techniques to deal with fault recovery and termination detection over the whole grid. Due to lack of space, we do not report the full algorithm, whose details can be found in [3].

2.2 The Extended Model

As stated above, in our extended model, *both* the data-graph *and* the RPQ are weighted by fuzzy quantities, to deal with the augmented modeling capabilities described above. The algorithm is generalized so as to perform a *semantic matching* of the information granules defined over the RPQ paths versus the knowledge stored in the DB. In other words, the RPQ is used not only to determine the set of acceptable paths, but also to compute on-the-fly the actual $c(\pi)$, by using a measure of *dissimilarity* between the DB and RPQ weights. Therefore, the novel model of the RPQ is a weighted FSA $A = (P, \Delta, \tau, \mathbb{K}, p_0, P_F)$, where $\mathbb{K}$ identifies the set of the edge weights. The transition relation τ is defined as $\tau \subseteq P \times \Delta \times \mathbb{K} \times P$, where the transition $t_i = (p_i, s_i, k_i, p'_i) \in \tau$ represents a query on the term labelled $s_i \in \Delta$ and weighted by

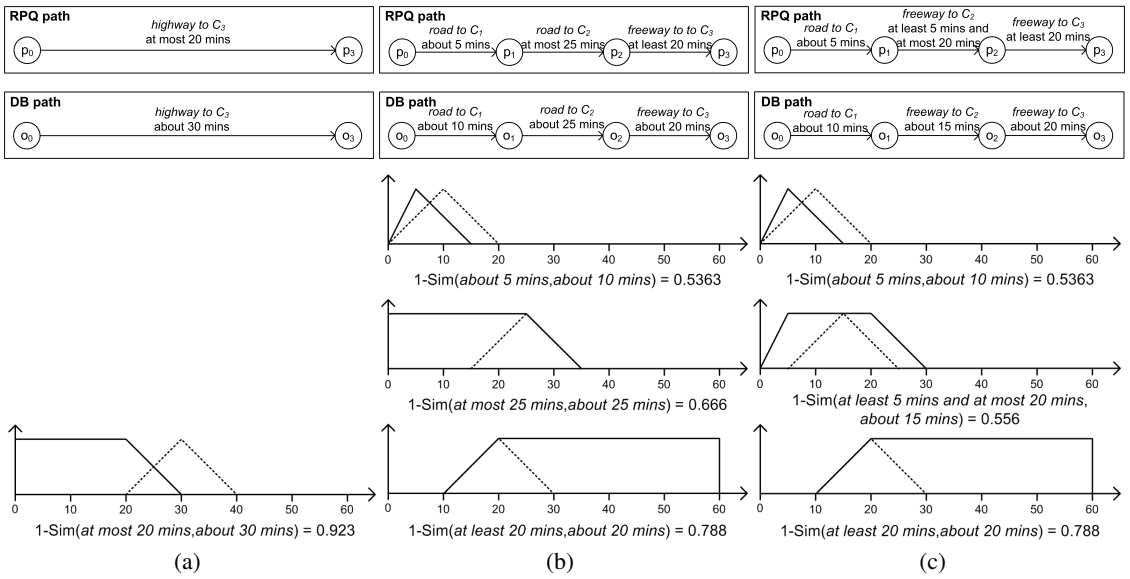

Fig. 3. RPQ paths and corresponding acceptable DB sub-graphs: (top) label matching and (bottom) fuzzy sets matching. Dotted lines represent DB weights whilst solid lines identify RPQ weights.

$k_i \in \mathbb{K}$. A simple example of RPQ weighted by fuzzy quantities is shown in Fig. 2(b). Obviously, the characteristics of the weight set $\mathbb{K}$ depend on the kind of information granulation that is chosen for knowledge representation. For this particular problem, $\mathbb{K}$ is a semi-ring whose elements are fuzzy quantities [8], with commutative sum $\oplus$ and distributive product $\otimes$ implementing the fuzzy union $\cup$ and intersection $\cap$ operators, respectively. We refer to this semi-ring as $\mathcal{F}$. In the light of this fuzzy interpretation, we associate each fuzzy quantity $k_i^l \in \mathcal{F}$ with a linguistic term t_i^l (e.g. $t_i^l = about\ 5\ mins$) by means of a mapping function $\mathcal{M} : \mathbb{T} \rightarrow \mathcal{F}$, such that $\mathcal{M}(t_i^l) = k_i^l$. We remark that the use of this association function allows to enforce the transparency and interpretability of the model by using meaningful linguistic terms in place of mathematical notations. In particular, the linguistic approach can be exploited to generate articulated fuzzy descriptions by applying linguistic modifiers [5] (such as "*at most*") to the primitive fuzzy sets, such as "*20 minutes*". Linguistic hedges have a clear mathematical formulation and act by modifying the shape of the fuzzy sets to which they are applied. Therefore it is possible for the user to reason using only linguistic terms and modifiers, thus hindering the complexity of the underlying mathematics.

To make the model clearer, consider the RPQ in Fig. 2(b): the weighted automaton A is described by the state set $P = \{p_0, p_1, p_2, p_3\}$, the signature $\Delta = \{$*road to* C_1,*freeway to* C_3, *highway to* $C_3\}$, the initial state p_0, the final state set $P_F = \{p_3\}$, and by the transition relation $\tau = \{(p_0,$ *road to* $C_1, k_1^1, p_1), \ldots, (p_2,$*freeway to* $C_3, k_3^1, p_3)\}$, where $\{k_1^1 = \mathcal{M}($*about 5 mins*$), \ldots, k_3^1 = \mathcal{M}($*at least 20 mins*$)\}$ defines the mapping between the linguistic terms and the fuzzy weights $k_i^l \in \mathcal{F}$.

Clearly, the three alternative paths defined by the RPQ in Fig. 2(b) over the spatial network of Fig. 2(a) are all acceptable. As stated above, from an information-granulation point-of-view, the challenge lies in discerning which of the three alternatives is the most similar to the request, i.e., in performing the semantic matching of the paths. To this aim, we exploit the notion of similarity of information granules that is common in the literature [9]. We use a similarity index to compute a dissimilarity measure $\overline{\mathrm{Sim}} : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{C}$ between each weight of the acceptable path on the DB and the corresponding weight of the RPQ. Each time a new partial path is discovered by the algorithm, the dissimilarity measure between the weights of the novel couple of matching edge is computed on-the-fly in the domain $\mathbb{C}$. The choice of the similarity index employed, and thus of $\mathbb{C}$, must be taken considering that we require certain properties both of $\oplus_{\mathbb{C}}$ and $\preceq_{\mathbb{C}}$. For instance, as in [6], we require that, given $a, b, c \in \mathbb{C}$, $a \preceq_{\mathbb{C}} b \Rightarrow (a \oplus_{\mathbb{C}} c) \preceq_{\mathbb{C}} (b \oplus_{\mathbb{C}} c)$.

Currently, our research is focused on formalizing the required properties and on defining a framework that should allow to derive a sound instance of the quintuple $\mathbb{Q} = (\mathbb{K}, \mathbb{C}, \oplus_{\mathbb{C}}, \preceq_{\mathbb{C}}, \overline{\mathrm{Sim}})$. At this stage, we are looking at two possible alternative strategies, restricted to the special case $\mathbb{K} \equiv \mathcal{F}$:

– *Early defuzzification*: we instantiate $\mathbb{Q}$ to $(\mathcal{F}, \mathbb{R}, \tilde{+}_{\mathbb{R}}, \leq_{\mathbb{R}}, 1 - \mathrm{Sim})$, where $\tilde{+}_{\mathbb{R}}$ is the average operation over reals and $\mathrm{Sim} : \mathcal{F} \times \mathcal{F} \rightarrow [0, 1]$ is a classical crisp evaluation of similarity of fuzzy quantities [9]. Implementing the cost sum

as the average operator $\tilde{+}_{\mathbb{R}}$ serves to unbias cost aggregation with respect to path length. Alternatively, if we intend to penalize longer paths as in [2, 3], we can define the $\oplus_{\mathbb{C}}$ operator as the real sum $+_{\mathbb{R}}$.

– *Late defuzzification*: we instantiate $\mathbb{Q}$ to $(\mathcal{F}, \mathcal{F}, \tilde{+}_{\mathcal{F}}, \leq_{\mathcal{F}}, \overline{\mathrm{Sim}})$, where $\tilde{+}_{\mathcal{F}}$ is the fuzzy extension of real operator $\tilde{+}_{\mathbb{R}}$, $\leq_{\mathcal{F}}$ is an ordering relation of fuzzy quantities [6, 7], and $\overline{\mathrm{Sim}} : \mathcal{F} \times \mathcal{F} \rightarrow \mathcal{F}_N$ is a fuzzy evaluation of dissimilarity of fuzzy quantities, with $\mathcal{F}_N$ defined as the set of fuzzy quantities defined over the real interval $[0, 1]$.

Whilst the early defuzzification option seems very easy to implement, it has the drawback of passing from the information granule representation to the crisp real representation too early, i.e., each time the on-the-fly cost of an edge is computed. On the other hand, the late defuzzification option is as powerful as challenging. Indeed, the possibility of representing $c(\pi)$ as a fuzzy quantity provides us with a much more significant evaluation of the actual similarity of the paths, but, on the other hand, implies a sound choice of $\leq_{\mathcal{F}}$ among the many alternative options in the literature, and the definition of a proper $\overline{\mathrm{Sim}}$.

As an example, let us consider the early defuzzification approach together with one of the similarity indices described in [9]. We define:

$$\overline{\mathrm{Sim}}(a, b) = 1 - \int_{x \in \mathbb{R}} (a \cap b) \,/\, \int_{x \in \mathbb{R}} (a \cup b)\,. \tag{1}$$

Consider the example in Fig. 3: the matching between the RPQ paths and the corresponding DB subgraphs (Fig. 3 (top)) can be evaluated by applying (1) to the couples of fuzzy weights depicted in Fig. 3 (bottom), obtaining $c(\pi_a) = 0.923$, $c(\pi_b) = 0.663$ and $c(\pi_c) = 0.627$ for each of the three acceptable paths in Fig. 3(a), 3(b) and 3(c), respectively. As expected, the cheapest DB path is the one that, by simple intuition, is linguistically most similar to the matching RPQ path.

3 Conclusion

The work is still in its preliminary phase, but opens several challenging theoretical issues. In particular, it would be interesting to study the general case of $\mathbb{K}$ as multi-dimensional space, that is $\mathbb{K} \subseteq \mathbb{K}_1 \times \mathbb{K}_2 \times ... \times \mathbb{K}_n$, where each $\mathbb{K}_i$ identifies a different domain knowledge, each represented by information granules of a (possibly) different type. Particular care must be taken to study the necessary conditions that need be satisfied by the domain-dependent dissimilarity functions, by a general cost sum $\oplus_{\mathbb{C}}$ and by the associated order relation $\preceq_{\mathbb{C}}$. Wang and Kerre offer an interesting starting point to tackle with this theoretical issues in their seminal works concerning advisable properties of the ordering of fuzzy quantities [6, 7]. Further, some work has to be done to explore the generation of a fuzzy model of similarity between fuzzy quantities that could be exploited in the late defuzzification approach. Indeed, it can be proved that, as far as we are concerned, the only proposal existing in the literature by Dubois and Prade [9, 10] cannot be used for fuzzy quantities ordering evaluation. On a more practical side, we intend to evaluate the performance of the model, and particularly of the two strategies described above, with respect to distributed data-graph querying, as done in [3].

References

1. Abiteboul, S., Buneman, P., Suciu, D.: Data on the Web: from relations to semistructured data and XML. Morgan Kaufmann, San Francisco, CA (1999)
2. Stefanescu, D.C., Thomo, A., Thomo, L.: Distributed evaluation of generalized path queries. In: Proceedings of the 2005 ACM Symposium on Applied computing (SAC'05), ACM Press (2005) 610–616
3. Miao, Z., Stefanescu, D.C., Thomo, A.: Grid-aware evaluation of regular path queries on spatial networks. In: Proceedings of the IEEE 21st International Conference on Advanced Information Networking and Applications (AINA07). (2007) (to appear)
4. Pedrycz, W.: Granular Computing - the emerging paradigm. Journal of Uncertain Systems **1**(1) (2007) 38–61
5. Klir, G.J., Yuan, B., eds.: Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lofti A. Zadeh. World Scientific Publishing, River Edge, NJ (1996)
6. Wang, X., Kerre, E.E.: Reasonable properties for the ordering of fuzzy quantities (i). Fuzzy Sets and Systems **118**(3) (2001) 375–385
7. Wang, X., Kerre, E.E.: Reasonable properties for the ordering of fuzzy quantities (ii). Fuzzy Sets and Systems **118**(3) (2001) 387–405
8. Bacciu, D., Botta, A., Melgratti, H.: A fuzzy approach for negotiating quality of services. In: Proceedings of the 2nd Symposium on Trustworthy Global Computing (TGC 2006). Volume 4661 of Lecture Notes in Computer Science., Springer (2007) 200–217
9. Cross, V.V., Sudkamp, T.A.: Similarity and compatibility in fuzzy set theory. Physica-Verlag, Heidelberg, Germany (2002)
10. Dubois, D., Prade, H.: A unifying view of comparison indices in a fuzzy set-theoretic framework. In Yager, R., ed.: Recent developments in fuzzy set and possibility theory. Pergamon Press, Elmsford, NY (1982) 3–13