"Same, Same but Different" A Survey on Duplicate Detection Methods for Situation Awareness*

Norbert Baumgartner[1], Wolfgang Gottesheim[2], Stefan Mitsch[2], Werner Retschitzegger[3], and Wieland Schwinger[2]

[1] team Communication Technology Mgt. Ltd., Goethegasse 3, 1010 Vienna, Austria
[2] Johannes Kepler University Linz, Altenbergerstr. 69, 4040 Linz, Austria
[3] University of Vienna, Dr.-Karl-Lueger-Ring 1, 1010 Vienna, Austria

Abstract. Systems supporting situation awareness typically deal with a vast stream of information about a large number of real-world objects anchored in time and space provided by multiple sources. These sources are often characterized by frequent updates, heterogeneous formats and most crucial, identical, incomplete and often even contradictory information. In this respect, duplicate detection methods are of paramount importance allowing to explore whether or not information having, e. g., different origins or different observation times concern one and the same real-world object. Although many such duplicate detection methods have been proposed in literature—each of them having different origins, pursuing different goals and often, by nature, being heavily domain-specific—the unique characteristics of situation awareness and their implications on the method's applicability were not the focus up to now. This paper examines existing duplicate detection methods appearing to be suitable in the area of situation awareness and identifies their strengths and shortcomings. As a prerequisite, based on a motivating case study in the domain of road traffic management, an evaluation framework is suggested, which categorizes the major requirements on duplicate detection methods with regard to situation awareness.

1 Introduction

Situation awareness. Situation awareness is gaining more and more importance as a way to cope with information overload in large-scale control systems, as e. g., encountered in the domain of road traffic management. As defined by Endsley [1], situation awareness comprises "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future", pursuing the goal of supporting human operators by pointing their attention to relevant sets of interrelated objects aggregated to situations (e. g., an accident causing a traffic jam). For this, systems supporting situation awareness typically deal with a

* This work has been funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under grant FIT-IT 819577.

vast stream of information about a large number of real-world objects anchored in time and space provided by multiple sources. These sources are often characterized by frequent updates, heterogeneous formats and most crucial, identical, incomplete, and often even contradictory information. Besides having to resolve structural heterogeneities at the *schema level*, the data itself has to be fused into a single consistent form at the *instance level* [2].

Duplicate detection. As a major prerequisite for the latter task, duplicate detection methods are of paramount importance allowing to explore whether or not information having, e. g., different origins, or different observation times concern one and the same real-world object. With appropriate duplicate detection methods, the number of entries describing the same real-world objects is reduced, thereby also increasing the *extensional conciseness* [2] of integrated data sources. To this end, a balance has to be found between the contrary goals of maximizing effectivity (i. e., finding all duplicates) and maximizing efficiency.

Duplicate detection for situation awareness. A series of duplicate detection methods has been already proposed for a wide range of application domains including, e. g., databases in general [3], temporal and geospatial databases in specific [4],[5], data warehouses [6], data stream management systems [7], sensor networks [8], XML data [9], and ontologies [10]—each of them pursuing different goals and often, by nature, being heavily domain-specific. Nevertheless, the unique characteristics of situation awareness, comprising fuzzy information about real-world objects anchored in time and space, object evolution, and context information, together with their implications on the methods' applicability were not the main focus up to now. Some of these characteristics are at least partially discussed in recent work, proposing e. g., a temporal similarity measure for finding comparable records in sequences [11], or dealing with the similarity of geospatial data [5]. Previous surveys in this realm [12],[13],[14],[15] however, have not yet comprehensively reviewed existing duplicate detection methods with regard to the specific characteristics of situation awareness.

This paper examines existing duplicate detection methods appearing to be suitable in the area of situation awareness and identifies their strengths and shortcomings. As a prerequisite, an evaluation framework is suggested, which categorizes the major requirements on duplicate detection methods with regard to the characteristics of situation awareness.

Structure of the paper. In the next section, we reflect on the domain of road traffic management to detail the characteristics of situation awareness, and thereby illustrate the requirements on duplicate detection. Section 3 proposes an evaluation framework for assessing duplicate detection methods with respect to their applicability for situation awareness. Based on this framework, a survey of selected duplicate detection methods follows in Section 4, resulting in several lessons learned and open issues for detecting duplicates in situation awareness in Section 5. Section 6 discusses related and complementary surveys, before we end the paper with a vision of future work in Section 7.

2 Road Traffic Management Case Study

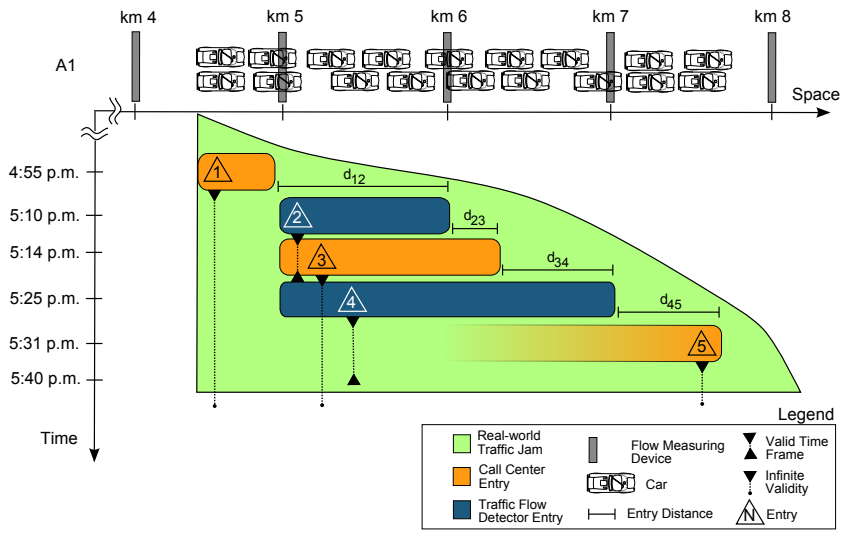
Road traffic management systems, being responsible for, e. g., improving traffic flow and ensuring safe driving conditions, are a typical application domain for situation awareness. Based on our experience in this area[4], examples from the domain of road traffic management are used to further illustrate the specific characteristics of situation awareness posing special requirements on duplicate detection. In principle, human operators of road traffic management systems observe highways for critical situations like traffic jams, relying on automated systems providing traffic information such as traffic flow detectors, but also on additional data sources like, e. g., motorists manually reporting traffic information to a call center.

Let us suppose a traffic jam builds up on a highway during rush hour, which may lead to a sequence of entries as depicted in Fig. 1a (1–5) in the road traffic management system, originating from various sources. Figure 1b shows how these entries and their different property values including categorical, temporal, and spatial properties are represented by the road traffic management system, assuming that structural heterogeneities between the different data sources have already been resolved.

From a chronological point of view, first of all a motorist, observing the traffic jam from the opposite lane while passing by, informs the call center (entry 1 in Table 1b). As the traffic jam's starting point is located, as depicted in Fig. 1a, between two traffic flow measuring devices, it takes a while until the traffic jam has grown to an extent also observed by the automated traffic flow detector (entry 2 in Table 1b), reporting updates every 15 minutes only (cf. property `validUntil` in Table 1b). As the traffic jam grows further, both the automated traffic flow detector and the call center *continue streaming* information about the traffic jam to the road traffic management system, as described by entries 3–5 in Table 1b. Motorists located at the end of the traffic jam are less and less able to observe it in its whole extent, resulting in inexact information about the traffic jam's starting point (entry 3), or even just in information about the traffic jam's end (entry 5).

Fuzzy information. Considering our scenario, first of all, duplicate detection methods have to deal with fuzzy information about real-world objects. Although the entries 1–4 describe the same traffic jam, they provide *contradictory* facts in the form of differing values for the properties `time`, `beginKm`, and `endKm`, as well as *incomplete* information (entry 5). Duplicate detection methods have to recognize that—despite this fuzzy information—all these entries concern the same real-world object. In this respect, uncertainty is unavoidable since entries can be compared using a similarity probability computed from their property values only. The challenge is to minimize uncertainty, even in the presence of such *temporal* and *spatial properties*, being represented not only in *quantitative*

[4] We are currently realizing ontology-driven situation awareness techniques for the domain of road traffic management together with our project partner Heusch/Boesefeldt, a German supplier of road traffic management systems.



(a) Example illustration.

	categorical		temporal		spatial		
Entry	type	source	time	validUntil	road	beginKm	endKm
1	Traffic Jam	Call center	4:55 p.m.	-	A1	4.4	4.7
2	Traffic Jam	Traffic Flow detector	5:10 p.m.	5:25 p.m.	A1	5	6
3	Traffic Jam	Call center	5:14 p.m.	-	A1	5	6.3
4	Traffic Jam	Traffic Flow detector	5:25 p.m.	5:40 p.m.	A1	5	7
5	Traffic Jam	Call center	5:31 p.m.	-	A1	?	7.5

(b) Entries with their properties as they are recorded in the data source.

Fig. 1: Information about a traffic jam during rush hour.

form (as in this example), but often also in *qualitative* form (e. g., in the form of spatial relations describing mereotopology and orientation). In this respect, temporality and spatiality are significant characteristics of situation awareness and should be dealt with independently from concrete application domains like road traffic management.

Object evolution. Besides fuzzy information, object evolution is an essential characteristic of situation awareness. As illustrated by our scenario, traffic information is not static over time, e. g., the traffic jam continuously grows in length, which is also reflected in the entries. But if these entries are compared on basis of their spatial and temporal properties only, one might conclude that entries 2 and 3 are probably duplicates and that entries 2 and 4 are also similar, but with less probability. All other entry pairs would show lower similarity, as depicted in Fig. 2. In particular, entry 1 most certainly would not match any of the other entries, and therefore it would remain undetected that entries 3 and 5 in fact update entry 1. Hence, to detect such duplicates it is necessary to take *object evolution*

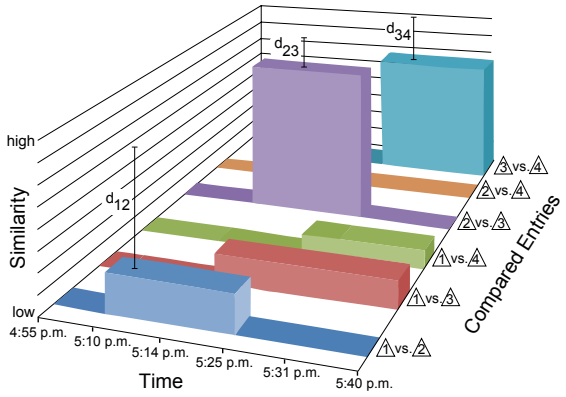


Fig. 2: Pairwise similarity of entries over time.

into account, and thereby reconstruct an *object's history* as a sequence of entries. Even if we were able to reconstruct object histories, the similarity probability between such object histories in different data sources would vary over time, as depicted in Fig. 3.

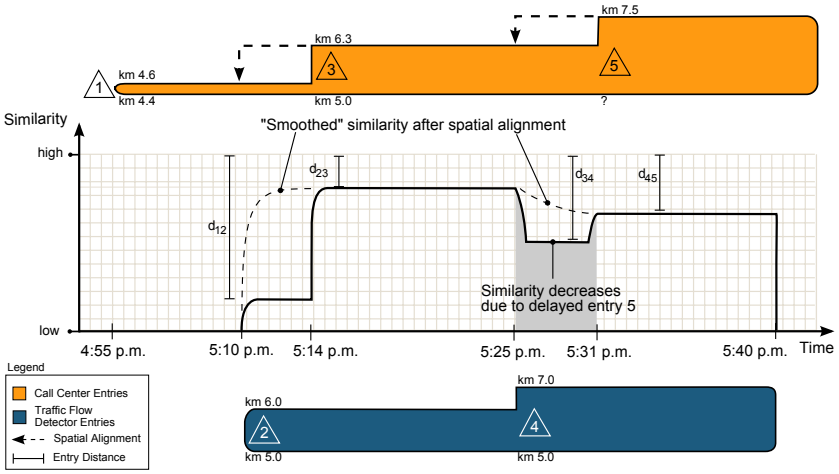


Fig. 3: Similarity of object histories in different data sources.

This variation in similarity probability can be partly accounted to fuzzy information, and partly stems from the fact that in the area of situation awareness different data sources report their updates in differing intervals. For example, our traffic flow detector reports updates constantly in equal-spaced intervals, whereas motorists report updates with varying intervals. Thus, before being able to meaningfully compare object histories from different data sources, in-

formation has to be *aligned*. In situation awareness, both *temporal alignments* and *spatial alignments* need to be supported. Temporal alignment associates entries according to the property `time` denoting when they were added to the data source. But, as highlighted in Fig. 3, in this example such an alignment leads to periods during which a new entry reported by the traffic flow detector is compared with an old—but still valid—entry reported by the call center. Hence, the reporting timestamps are not always adequate for matching entries between these two data sources. Alternatively, spatial alignment, using the similarity of spatial properties, can be performed. In our example, we may use the similarity of traffic jam lengths and positions as a way to identify matching entries, as also illustrated in Fig. 3. Such an alignment reveals that the call center reports updates with a delay of four to six minutes, compared to the traffic flow detector, and thereby facilitates duplicate detection.

Context information. The specific characteristics of situation awareness as discussed above requires duplicate detection methods to consider additional information not contained within entries in terms of *context information* [16]. Context information is vital to accurately interpret entries as it provides details on a data source's environment. For example, the distance between the `beginKm` properties of entry 1 and 2 (600m) can only be interpreted correctly when taking into account *spatial granularity* describing the denseness of the road network itself: on a highway, such a distance is rather low, whereas in urban areas with a dense road network the same distance describes substantially different positions. Additionally, in case a growing traffic jam, starting e. g. on highway "A1", evolves onto a different highway, information about *spatial topology* describing the road network's layout must be considered. Similar characteristics are encountered in the temporal dimension: in order to correctly interpret differences between entries reported by the traffic flow detector, we need to take into account the traffic flow detector's 15-minute update interval referred to as *temporal granularity*.

Besides context information about spatial and temporal properties, context information about object evolution should be considered. The entry sequence 1-3-5 provided by the call center, e. g., describes the typical phases of a traffic jam: most traffic jams build up by growing at their end, then move as cars at the beginning leave the jam at the same pace as cars are joining at the end, and finally shrink if the cars leaving at the beginning outnumber the cars joining at the end. Such *evolution patterns* describing the typical behavior of objects facilitate the reconstruction of object histories.

Human decision making. As some kind of cross-cutting characteristic with respect to the previous ones, situation-aware systems most often support human operators required to make decisions and take actions having impact on the real world: for example, they need to issue warnings of traffic jams being ahead to motorists, in order to pro-actively prevent critical situations. If such warnings are omitted due to false results of duplicate detection methods, serious consequences, like accidents, are to be expected. Therefore, the results of duplicate detection—which are not always obvious to human operators—should be *allegeable* and *traceable*, in order to allow operators to question these results.

Summary. Summarizing, the characteristics of situation awareness, as illustrated in the examples above, lead to the following requirements on duplicate detection: First, fuzzyness in terms of contradictory and incomplete information about real-world objects being continuously streamed require to deal with temporal and spatial properties of quantitative and qualitative nature in a performant way. Second, evolution comprising changes in position and extent of objects as well as different update intervals and delays between observation and reporting demands for a reconstruction and alignment of object histories. Third, the meaning of property values dependending on the particular environment they are obtained from makes it necessary to consider context information in the form of spatial and temporal granularity, spatial and temporal topology, as well as evolution patterns. Finally, the impacts that actions—taken by human operators in response to the results of situation-aware systems—may have on the real world, demand for allegeable and traceable results, in order to increase the confidence of human operators about the system.

3 Evaluation Framework

Based on the requirements laid out in the previous section, we propose an evaluation framework for duplicate detection methods. We provide ten criteria for measuring the applicability of duplicate detection methods to situation awareness and categorize them into three orthogonal dimensions—fuzzy information, object evolution, and context information—with human decision making crosscutting the other dimensions, as depicted in Fig. 4.

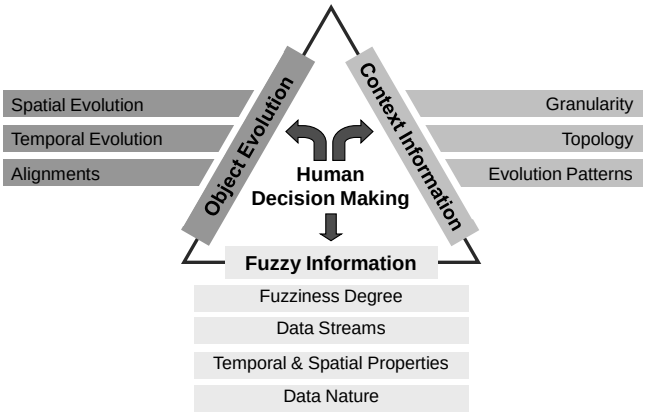


Fig. 4: Overview of the evaluation framework.

In the following, the selected criteria are described in detail comprising a name, a definition, an indication if it was already defined in previous work, and measurable parameters.

Fuzzy information. This category considers criteria measuring the degree to which duplicate detection methods meet the requirements stemming from fuzzy information about real-world objects.

Fuzziness Degree. As laid out in the examples above, besides identical information, duplicate detection methods have to deal with fuzziness in the form of contradictory and incomplete information. We therefore distinguish methods according to the fuzziness degree (ranging from *identical* to *contradictory* and *incomplete*) they are able to handle.

Data Streams. Considering the requirements of data streams about real-world objects prevalent in situation awareness, we distinguish duplicate detection methods according to whether or not they support such data streams.

Temporal and Spatial Properties. For temporal and spatial properties, as discussed in the examples above, we need similarity functions measuring the similarity between two property values. This criterion is an extension of the *field-matching techniques* distinction in [13], which takes into account string-based and numeric metrics only. We distinguish duplicate detection methods according to their support of *temporal* and *spatial* properties.

Data Nature. As motivated in the example, duplicate detection methods have to support quantitative data, e. g., spatial positions or timestamps according to global reference systems like WGS-84 [17] or UTC [18], as well as qualitative data, like road names or spatial and temporal relations taking into account mereotopology and orientation [19]. Therefore, we distinguish duplicate detection methods with respect to the supported nature of data, which can be *quantitative* and/or *qualitative*.

Object evolution. This category contains criteria measuring the extent to which duplicate detection methods are able to consider object evolution.

Spatial Evolution. In situation awareness, objects evolve in space by changing their position as well as their extent (defined in [20] as change in topology, and change in geometry). We distinguish duplicate detection methods into those supporting *both kinds* of evolutions, such methods only supporting changes of an object's *position* or an object's *size*, and those not supporting spatial evolution.

Temporal Evolution. Temporal evolutions of real-world objects make it necessary to reconstruct object histories from entries forming sequences in a data source, as the examples above illustrated with consecutive reports on changing traffic jam positions and lengths. We distinguish between methods supporting the reconstruction of object histories, and others that do not.

Alignments. In order to meaningfully compare object histories, they first need to be aligned, i. e. correspondences between the entries in different histories need to be established using, e. g., timestamps or spatial information. We distinguish between duplicate detection methods supporting *temporal alignment*, *spatial alignment* [21], and/or *other* forms of alignments, and those not being able to align object histories. All such alignment forms can be additionally subdivided by the fact whether or not they are able to align sequences of different lengths [22].

Context information. The criteria in this category measure the extent to which duplicate detection methods are able to exploit context information.

Granularity. The granularity of properties is vital for duplicate detection methods to compare property values in situation awareness. We distinguish between duplicate detection methods interpreting *temporal* [23], *spatial* [24],[25], and/or granularity of *other* properties, and such methods not supporting granularity.
Topology. We distinguish between duplicate detection methods interpreting *temporal*, *spatial*, and/or *other kinds* of topologies, and those not supporting any topologies.
Evolution Patterns. Real-world objects in situation awareness change frequently, thereby often following known evolution patterns. Duplicate detection methods supporting evolution patterns are likely to yield better effectivity in situation awareness, in comparison to methods that do not. This criterion bases on the evolution of situations, as we have proposed in [26].
Human decision making. Human decision making bases on explanations, as measured by the following criterion.
Explanations. To increase acceptance amongst human operators, as indicated in the example, the decision whether two objects are duplicates or not must be allegeable and traceable. Methods based on *logic* being able to give explanations themselves by retracing their inference steps best suit situation awareness, whereas the results of *deterministic* methods, like decision trees or rules, can at least be comprehended by domain experts, and least suited are *non-deterministic* methods (e. g., support vector machines or neural networks) not providing explanations.

4 Survey of Duplicate Detection Methods

According to the evaluation framework introduced above, we examine existing duplicate detection methods appearing to be suitable in the area of situation awareness and identify their strengths and shortcomings. The methods in this survey originate from a wide range of application domains including databases in general, temporal and geospatial databases in specific, data warehouses, data stream management systems, sensor networks, XML data, ontologies, and moving object trajectories. In the following, domains with similar approaches to duplicate detection are combined into groups and presented in the order of increasing applicability to situation awareness. In each such group, structured along the evaluation framework's four dimensions, we informally discuss how a representative approach meets our criteria proposed above.
Data stream management systems. Data stream management systems are designed for accessing data streams, as e. g., encountered in Internet advertisement click streams [27], with database-like query interfaces [28]. In such data stream management streams, duplicate detection is vital, e. g., to detect frauds or to analyze trends. In this respect, several works on duplicate detection exist [7], [27]. As these works approach duplicate detection with rather similar methods, we evaluate the work of Metwally [27] being one of the pioneers in this domain. Metwally focuses on searching duplicates in a single pass with performant algorithms. As a consequence, several simplifications accross all criteria dimensions

are assumed, such as objects having unique identifiers therefore not supporting *fuzzy information* besides data streams. Moreover, neither *object evolution* nor *context information* and *human decision making* is considered. Nevertheless, if the prerequisite of generating unique identifiers is solved with methods from other domains, we can still learn how to handle the volatility of data streams during duplicate detection.

Temporal databases. Temporal databases, which store additional timestamps for describing an entry's validity and input time [29], describe temporal evolution of objects by multiple entries having adjacent valid times. To the best of our knowledge, in such temporal databases, like described in [30] and [4], duplicate detection is simplified to enable eliminating identical entries originating, e. g., from join-operators. We evaluate Slivinskas et al. [4] explicitly proposing a temporal relational algebra handling such duplicate entries being characterized by identical properties with overlapping validity. However, besides temporality, *fuzzy information*, *object evolution*, as well as *context information* are not discussed. Hence, except for the representation of temporal evolution, the method of Slivinskas et al. is less suitable for supporting situation awareness.

Sensor networks. In the area of sensor networks, we encounter a rather opposite problem to duplicate detection referred to as outlier detection, aiming to make values on property-level more robust against noise or failing sensors [31]. Nevertheless, methods proposed in this area, like [32], and [8], are useful for our purposes, because many situation-aware systems in fact base on sensor networks for observing real-world objects. Of such methods, Jefferey's sensor data cleaning approach [8] appears to be most suitable to situation awareness due to the exploitation of spatial and temporal characteristics of sensor data. Operating on data streams, Jefferey counteracts *fuzzy information* (describing quantitative values, e. g., temperature) by "smoothing" them using *context information* like spatially and temporally nearby values determined on basis of the sensor network's spatial and temporal topology. Thereby, the particular smoothing function is determined by likely evolution patterns of the observed properties. Additionally, segments in time and space, in which values are expected to be homogeneous, define the method's temporal and spatial granularity. These granularities remain, however, stable over time, because *object evolution* is not considered.

Databases, data warehouses, and XML data. In databases, data warehouses and XML data, duplicate detection is most often a prerequisite to fusion of data [2]. In this respect, duplicate detection is performed typically in an offline manner (either during nightly cleaning runs in databases, or during loading of data warehouses), and methods are characterized by configurable processes based on generic similarity measures. Numerous approaches exist, like [9], [33], [34], and [6]. We evaluate DogmatiX [9]—originally developed for XML and later adopted to relational databases—due to its proven applicability in real-world data cleaning projects [35], and its additional features comprising independence from data models (e. g., relational or XML) and heuristics supporting domain experts during configuration. DogmatiX handles *fuzzy information* by compar-

ing quantitative string and numeric properties in static data sources. As the major application areas of DogmatiX, like credit record databases, do not focus on tracking objects in time and space, *object evolution* is not supported. DogmatiX uses *context information*, like the granularity of numeric properties, during configuration only. But due to the absence of spatial and temporal similarity functions, neither spatial/temporal granularity nor topology are exploited. The results of DogmatiX base on deterministic rules implemented in its classifiers, thereby supporting *human decision making* with limited explanations only.

Ontologies. Ontologies have recently been regarded to be beneficial for achieving situation awareness, because of their semantically-rich kind of information representation often being based on qualitative data. Numerous works exist in the field of ontology mapping [36],[37], but actually only a small fraction of them, like [38] and [39], discuss ontology-matching on the instance-level. As a representative for such methods, we evaluate the work of Qin [39], which explicitly features duplicate detection based on the method proposed by Dong et al. [16]. This method detects duplicates in the presence of *fuzzy information* based on string and numeric properties. The proposed duplicate detection method, however, assumes strictly monotonic-raising property similarity values in order for the propagation algorithm to terminate, and hence, it is not applicable to data streams. Moreover, the method does not consider *object evolution*. *Context information* in terms of relations between objects is used to structure likely duplicates into a dependency graph. The work's main contribution is a similarity propagation algorithm using this dependency graph to revisit and re-evaluate neighbors of detected duplicates (e. g., if two publications are accounted to be equal, their authors are very likely also duplicates). The nature of this similarity propagation algorithm built on graph-structures is deterministic, but the method does not exploit the full potential of ontologies to support *human decision making* with automatically inferred explanations.

Geospatial databases. In geospatial database research, duplicate detection is seen as a part of information integration combining multiple heterogeneous sources [40]. Among existing approaches (e. g., [41], [42], and [43]) we evaluate the work of Bakillah et al. [43] due to their notion of spatial evolution taking into account changes in extent. The proposed method handles *fuzzy information* with similarity measures for spatial on-line analytical processing cubes (spatial OLAP—SOLAP), being an extension to OLAP cubes used in data warehouses [44]. Additionally, Bakillah et al. consider *object evolution* in time and space, whereby spatial evolution is supported in terms of extent only. In order to determine weights for combining similarity measures, the method comprises a semantic similarity model based on an ontology taking user-defined *context information* into account. Overall similarity is calculated by a deterministic formula, thereby supporting *human decision making* in a limited way only.

Moving objects trajectories. Similarity analysis of trajectories is an area concerned with comparing the traces of moving objects in time and space [45]. Several such methods exist, like [46], [47], and [48]. However, most of them in fact measure similarity in Euclidian space only. We evaluate the work of Hwang et al.

[48] due to its applicability for describing road networks with an alternative spatial representation. Hwang et al. propose temporal, spatial, and spatio-temporal similarity measures to counteract *fuzzy information* in trajectories of moving objects. Being based on qualitative spatial and temporal information in terms of "points of interest" and "times of interest", such object trajectories, however, must be fully constructed before comparison can take place. Hence, the method does not support data streams. Moreover, this implies that the method can in fact handle identical and contradictory information only, but assumes trajectories to be complete. *Object evolution* in the form of position changes, as well as alignments by temporal and spatial similarity are possible. As a prerequisite, *context information* describing spatial and temporal granularity, as well as spatial topology of road networks must be available to the method. The proposed similarity measures are defined using deterministic functions, allowing limited support of *human decision making* only.

5 Lessons Learned

Our survey in the previous section revealed that none of the investigated duplicate detection methods originating from various areas fulfills all criteria of situation awareness, as summarized in Table 1, but at least concepts for specific subproblems are proposed. In the following, we point out lessons learned to highlight open research questions.

	Fuzzy Information				Object Evolution			Context Info.			
	Fuzziness Degree	Data Streams	Temporal and Spatial Properties	Data Nature	Spatial Evolution	Temporal Evolution	Alignments	Granularity	Topology	Evolution Patterns	Explanations
Data Stream Mgt. Systems (Metwally et al.)	–	✓	–	QN	–	–	–	–	–	–	–
Temporal Databases (Slivinskas et al.)	–	–	T	QN	–	~	–	–	–	–	–
Sensor Networks (Jefferey et al.)	–	✓	–	QN	–	–	–	S,T	S,T	~	–
(Relational) Databases (Weis et al.)	✓	–	–	QN	–	–	–	O	–	–	~
Ontologies (Qin et al.)	✓	–	–	QN	–	–	–	–	O	–	~
Geospatial Databases (Bakillah et al.)	✓	–	S,T	QN	E	~	–	~	O	–	~
Moving Object Trajectories (Hwang et al.)	~	–	S,T	QL	P	–	S,T	S,T	S	–	~

Legend

✓ Fully Supported	S Spatial	P Position	QN Quantitative
~ Partially Supported	T Temporal	E Extent	QL Qualitative
– Not Supported	O Other		

Table 1: Survey summary.

Data streams not considered in presence of fuzzy information. For data stream support, one can observe that, with two exceptions (the areas of data streams and sensor networks), the surveyed methods detect duplicates in an off-line manner only. In particular, duplicate detection in data streams appears to be dependent on the existence of unique object identifiers, not least due to performance requirements posed on such methods by high data volumes.

Spatial and temporal properties supported, but similarity measures in their infancy. Spatial and temporal properties, represented in various formats, are supported by different domains. However, similarity measures for such properties are still in their infancy: In the surveyed temporal database methods, temporal similarity is defined simply in terms of overlapping time periods (independent from their lengths, or the amount to which they overlap), whereas in the trajectories group similarity is defined in terms of equality of points and times of interest. At least in the domain of geospatial databases, being often based on quantitative data, different functions measuring spatial similarity are proposed.

Qualitative nature of data not exploited. Most methods, as can be seen in Table 1 work on quantitative data, in order to facilitate the computation of similarity probabilities. In the presence of qualitative data, computing such similarity probabilities is often a challenging task demanding domain knowledge. At least, one group of methods (namely, from the area of trajectories) supports qualitative spatial and temporal properties. Their notion of points of interests and times of interests, however, is of a rather informal nature and tailored to their application domain. In duplicate detection, formal specifications of qualitative data already introduced in situation awareness [49], are still missing, leading to incompatible methods being only applicable in their application domain.

Object evolution support not an issue. Current approaches largely fail to support object evolution. Spatial evolution is discussed in the group of trajectory methods and in geospatial databases, with both groups only supporting a part of the requirements of situation awareness. Spatial alignments are only offered by the group of trajectory methods, whereby these alignments base on the same rather basic similarity measures already described above. Temporal evolution is partially supported in temporal databases, which regard entries having identical property values and adjacent time periods to represent a temporal evolution, as well as in geospatial databases. The combination of both spatial and temporal evolution, which means reconstructing object histories in presence of moving objects additionally evolving in size, still needs to be solved.

Context information is widely adopted, but evolution patterns are not exploited. Spatial, temporal and other forms of granularity, as well as topology context information is present in various domains, but except for the group of sensor networks, evolution patterns are not incorporated into duplicate detection methods. Such sensor networks, however, rather use evolution patterns to select appropriate algorithms during implementation (e. g., room temperature can only evolve steadily, allowing to remove implausible values with median functions),

and not, as envisioned, during runtime as additional input to duplicate detection methods used for reconstructing object histories.

Representation of context information is domain-specific. In general, context information is regarded by the surveyed methods as domain-specific knowledge, and hence, no effort is put into making context information representations domain-independent. Moreover, often such context information is even tight-knit with algorithm implementations, resulting in methods not being applicable outside their domain.

Automated inference of explanations is not considered. The results of the evaluated methods, all being based on deterministic algorithms, can at least be comprehended by domain experts. However, automatically giving explanations to human operators is still an issue to be solved. In this respect, we regard duplicate detection based on reasoning with ontologies to be superior to existing approaches. However, as the survey revealed, duplicate detection in ontologies is still in its infancy, making it necessary to incorporate concepts from other domains.

6 Related Surveys

As indicated in the previous section, duplicate detection is a major issue in a wide range of domains and such methods, therefore, have been already compared in previous surveys. This section outlines related surveys with respect to their discussion of the characteristics of situation awareness, starting with the prominent area of duplicate detection in databases in general, moving on to knowledge discovery in temporal and geospatial databases in specific, and finishing with surveys in the area of qualitative data.

The surveys of Bleiholder and Naumann [2], Elmagarmid et al. [13], and Herzog et al. [12] recognize duplicate detection in databases as a highly domain- and even dataset-specific task. In [2], duplicate detection is discussed in the larger context of information integration, as the intermediate step between schema mapping and data fusion. The authors emphasize the need for effective and efficient property similarity functions that are able to operate on large datasets, but as the focus of the survey is on data fusion, duplicate detection is not further elaborated. Elmagarmid et al. [13] survey approaches and algorithms for property similarity functions in terms of attribute matching and record matching, and for improving the efficiency of duplicate detection in databases. As the authors consider attribute matching to be a string similarity problem, neither the specifics of comparing temporal nor those of spatial data are taken into account. In a similar survey, [12] provides an extensive overview on probabilistic algorithms for entry similarity computation with a strong focus on the identification of persons and addresses in large datasets. All three surveys, however, do neither discuss property similarity functions for temporal and spatial data, nor object evolution and context information, and are, therefore, less suitable for situation awareness.

Methods dealing with temporal data are subject of a survey on temporal knowledge discovery [50]. According to the classification scheme of temporal data presented in this survey, in situation awareness we have to deal with the most complex category of *fully temporal* information comprising sequences of time-stamped data, as can also be seen in the case study above. In order to evaluate knowledge discovery methods dealing with such information, the authors present a taxonomy structuring temporal knowledge discovery into methods based on a-priori knowledge (like sequence mining methods), and such based on machine learning. The survey, however, focuses on methods discovering knowledge from entry sequences. As a consequence, neither property similarity functions nor context information are discussed. A similar research field is discussed in a survey of clustering methods for time series data [22]. Most interesting for our work is the discussion of similarity functions for comparing time series, i. e., entry sequences, which, however, focuses on quantitative data only.

The survey of Schwering [14] focuses on semantic similarity of geospatial data, thereby emphasizing the importance of different spatial representation models for interpreting semantic similarity, which we subsume under the term of spatial context information. Other characteristics of situation awareness, besides spatial similarity, are not discussed in this survey.

Whereas the surveys discussed above strongly focus on quantitative data, a complementary survey on similarity functions based on qualitative data, particularly categorical data, is presented in [51]. The survey, however, discusses mainly performance characteristics of such functions and, to this end, uses the same function for comparing all properties of different entries. In their concluding remarks, the authors highlight the importance of similarity functions being tailored to the characteristics of single properties. We argue, that in situation awareness we additionally need to incorporate context information to account for spatial and temporal variations of a single property when computing similarity.

7 Future Work

Further research questions arise from the lessons learned discussed above and include the exploitation of formal specifications of qualitative spatial and temporal properties for measuring similarity in data streams, as well as the integration of context information to better support object evolution during duplicate detection in situation awareness. In this respect, we argue that, based on our previous work [49], ontologies could be beneficial as a specification formalism for such qualitative properties, as well as for context information. Thereby, similarity of entries can not only be assessed using existing reasoning techniques for knowledge inference, but can also be seen in the broader context of reasoning about situations in situation awareness. In particular, we aim to exploit our notion of *neighborhood of situations* introduced in [26]. For this, we need to develop appropriate functions measuring the distance between situations in our ontology. Such functions can partly base, on the one hand, on concepts for describing temporal and spatial similarity of qualitative information proposed in the field of moving

object trajectories, and on the other hand, on methods exploiting temporal and spatial evolution, as well as context information. Considering object evolution, however, evolution patterns are not taken into account by the surveyed duplicate detection methods. Filling this gap will, therefore, be of special interest in future research: an ontology for describing evolution patterns, as well as similarity functions exploiting these patterns still need to be developed. As an additional bonus, such an ontology-driven system enables automated inference of explanations. We will further investigate the envisioned approach with respect to its applicability in a real-world scenario in the scope of our ongoing research project BeAware!, which focuses on ontology-driven situation awareness.

References

1. Endsley, M.R.: Design and evaluation for situation awareness enhancement. In: Proceedings of the Human Factors Society 32nd Annual Meeting, Santa Monica, CA, USA, Human Factors Society (1988) 97–101
2. Bleiholder, J., Naumann, F.: Data fusion. ACM Computing Surveys **41**(1) (2008)
3. Hernández, M.A., Stolfo, S.J.: Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery **2**(1) (1998) 9–37
4. Slivinskas, G., Jensen, C.S., Snodgrass, R.T.: A foundation for conventional and temporal query optimization addressing duplicates and ordering. IEEE Transactions on Knowledge and Data Engineering **13**(1) (2001) 21–49
5. Schwering, A., Raubal, M.: Measuring semantic similarity between geospatial conceptual regions. In: Proceedings of the 1st International Conference on GeoSpatial Semantics, Mexico City, Mexico (2005) 90–106
6. Ananthakrishna, R., Chaudhuri, S., Ganti, V.: Eliminating fuzzy duplicates in data warehouses. In: Proceedings of the 28th International Conference on Very Large Data Bases, VLDB Endowment (2002) 586–597
7. Deng, F., Rafiei, D.: Approximately detecting duplicates for streaming data using stable bloom filters. In: Proceedings of the 2006 ACM SIGMOD Intl. Conference on Management of Data, New York, NY, USA, ACM Press (2006) 25–36
8. Jefferey, S.R., Alonso, G., Franklin, M.J., Hong, W., Widom, J.: Declarative support for sensor data cleaning. In: Proceedings of the 4th International Conference on Pervasive Computing, Dublin, Ireland, Springer (2006) 83–100
9. Weis, M., Naumann, F.: Dogmatix Tracks Down Duplicates in XML. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, ACM Press (June 2005) 431–442
10. Noy, N.F.: Semantic integration: A survey of ontology-based approaches. SIGMOD Rec. **33**(4) (2004) 65–70
11. Wongsuphasawat, K., Shneiderman, B.: Finding comparable temporal categorical records: A similarity measure with an interactive visualization. Technical Report HCIL-2009-08, University of Maryland (2009)
12. Herzog, T.N., Scheuren, F.J., Winkler, W.E.: Data Quality and Record Linkage Techniques. Springer (2007)
13. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering **19**(1) (2007) 1–16
14. Schwering, A.: Approaches to semantic similarity measurement for geo-spatial data: A survey. Transactions in GIS **12**(1) (2008) 5–29

15. Morris, A., Velegrakis, Y., Bouquet, P.: Entity identification on the semantic web. In: Proceedings of the 5th International Workshop on Semantic Web Applications and Perspectives, Rome, Italy (2008)
16. Dong, X., Halevy, A., Madhavan, J.: Reference reconciliation in complex information spaces. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, ACM Press (2005) 85–96
17. Mularie, W.M.: World Geodetic System 1984–Its Definition and Relationships with Local Geodetic Systems. Technical Report TR8350.2, National Imagery and Mapping Agency (2000)
18. ITU-R: TF.460-4, Annex I. International Telecommunication Union (1970)
19. Baumgartner, N., Retschitzegger, W.: Towards a situation awareness framework based on primitive relations. In: Proceedings of the IEEE Conference on Information, Decision, and Control (IDC), Adelaide, Australia, IEEE (2007) 291–295
20. Abraham, T., Roddick, J.F.: Survey of spatio-temporal databases. GeoInformatica **3**(1) (1999) 61–99
21. Caspi, Y., Irani, M.: Spatio-temporal alignment of sequences. IEEE Transactions on Pattern Analysis Machine Intelligence **24**(11) (2002) 1409–1424
22. Liao, T.W.: Clustering of time series data–a survey. Pattern Recognition **38**(11) (2005) 1857–1874
23. Dyreson, C.E., Evans, W., Lin, H., Snodgrass, R.T.: Efficiently supporting temporal granularities. IEEE Trans. on Knowledge and Data Eng. **12**(4) (2000) 568–587
24. Worboys, M.: Computation with imprecise geospatial data. Computer, Environment and Urban Systems **22**(2) (1998) 85–106
25. Khatri, V., Ram, S., Snodgrass, R.T., O'Brien, G.M.: Supporting user-defined granularities in a spatiotemporal conceptual model. Annals of Mathematics and Artificial Intelligence **36**(1-2) (2002) 195–232
26. Baumgartner, N., Retschitzegger, W., Schwinger, W., Kotsis, G., Schwietering, C.: Of situations and their neighbors—Evolution and Similarity in Ontology-Based Approaches to Situation Awareness. In: Proceedings of the 6th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT), Roskilde, Denmark, Springer (2007) 29–42
27. Metwally, A., Agrawal, D., El Abbadi, A.: Duplicate detection in click streams. In: Proceedings of the 14th International Conference on World Wide Web, New York, NY, USA, ACM (2005) 12–21
28. Cheng, J., Ke, Y., Ng, W.: A survey on algorithms for mining frequent itemsets over data streams. Knowledge and Information Systems **16**(1) (2008) 1–27
29. Jensen, C.S., Snodgrass, R.T.: Temporal data management. IEEE Transactions on Knowledge and Data Engineering **11**(1) (1999) 36–44
30. Dekhtyar, A., Ross, R., Subrahmanian, V.S.: Probabilistic temporal databases, I: Algebra. ACM Transactions on Database Systems **26**(1) (2001) 41–95
31. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor network survey. Computer Networks **52**(12) (2008) 2292–2330
32. Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., Gunopulos, D.: Online outlier detection in sensor data using non-parametric models. In: Proc. of the 32nd Intl. Conf. on Very Large Data Bases, VLDB Endowment (2006) 187–198
33. Thor, A., Rahm, E.: MOMA - A Mapping-based Object Matching System. In: Proc. of the 3rd Biennial Conf. on Innovative Data Systems Research, Asilomar, CA, USA (2007) 247–258
34. Rusu, L.I., Rahayu, J.W., Taniar, D.: On data cleaning in building XML data warehouses. In: Proceedings of the 6th International Conference on Information Integra-

tion and Web-based Applications Services, Jakarta, Indonesia, Austrian Computer Society (2004)
35. Weis, M., Naumann, F., Jehle, U., Lufter, J., Schuster, H.: Industry-scale duplicate detection. Proceedings of the VLDB Endowment **1**(2) (2008) 1253–1264
36. Kalfoglou, Y., Schorlemmer, M.: Ontology Mapping: The State of the Art. The Knowledge Engineering Review **18**(1) (2003) 1–31
37. Choi, N., Song, I.Y., Han, H.: A survey on ontology mapping. ACM SIGMOD Record **35**(3) (2006) 34–41
38. Castano, S., Ferrara, A., Lorusso, D., Montanelli, S.: On the ontology instance matching problem. In: Proceedings of the 19th International Conference on Database and Expert Systems Applications, Turin, Italy, IEEE (2008) 180–184
39. Qin, H., Dou, D., LePendu, P.: Discovering executable semantic mappings between ontologies. In: Proc. of the OTM Confederated Intl. Conf. CoopIS, DOA, ODBASE, GADA, and IS 2007, Vilamoura, Portugal, Springer (2007) 832–849
40. Beeri, C., Kanza, Y., Safra, E., Sagiv, Y.: Object fusion in geographic information systems. In: Proceedings of the Thirtieth international conference on Very Large Data Bases, VLDB Endowment (2004) 816–827
41. Sehgal, V., Getoor, L., Viechnicki, P.D.: Entity resolution in geospatial data integration. In: Proc. of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems, New York, NY, USA, ACM Press (2006) 83–90
42. Rodríguez, M.A., Bertossi, L., Caniupán, M.: An inconsistency tolerant approach to querying spatial databases. In: Proc. of the 16th Intl. Conf. on Advances in Geographic Information Systems, New York, NY, USA, ACM Press (2008) 1–10
43. Bakillah, M., Mostafavi, M.A., Bédard, Y.: A semantic similarity model for mapping between evolving geospatial data cubes. In: LNCS. Volume 4278/2006. (2006)
44. Rivest, S., Bdard, Y., Proulx, M.J., Nadeau, M., Hubert, F., Pastor, J.: SOLAP technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. ISPRS Journal of Photogrammetry and Remote Sensing **60**(1) (2005) 17 – 33
45. Frentzos, E., Pelekis, N., Ntoutsi, I., Theodoridis, Y.: Trajectory Database Systems. In: Mobility, Data Mining and Privacy—Geographic Knowledge Discovery. Springer (2008) 151–188
46. Chen, L., Özsu, M.T., Oria, V.: Robust and fast similarity search for moving object trajectories. In: Proceedings of the International Conference on Management of Data, New York, NY, USA, ACM (2005) 491–502
47. Frentzos, E., Gratsias, K., Theodoridis, Y.: Index-based most similar trajectory search. In: Proc. of the 23rd Int. Conf. on Data Engineering, IEEE (2007) 816–825
48. Hwang, J.R., Kang, H.Y., Li, K.J.: Searching for similar trajectories on road networks using spatio-temporal similarity. In: Proc. of the 10th East Euro. Conf. on Adv. in Databases and Inf. Sys., Thessaloniki, Greece, Springer (2006) 282–295
49. Baumgartner, N., Retschitzegger, W., Schwinger, W.: Lost in time, space, and meaning—an ontology-based approach to road traffic situation awareness. In: Proc. of the 3rd Worksh. on Context Awareness for Proactive Sys., Guildford, UK (2007)
50. Roddick, J.F., Spiliopoulou, M.: A survey of temporal knowledge discovery paradigms and methods. IEEE Trans. on Knowl. and Data Eng. **14**(4) (2002)
51. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the SIAM International Conference on Data Mining, Atlanta, GA, USA, SIAM (2008) 243–254