Information Fusion using Conceptual Graphs: a TV Programs Case Study

Claire Laudy[1,2] and Jean-Gabriel Ganascia[2]

[1] THALES Research & Technology, Palaiseau, France
[2] ACASA, Laboratoire d'Informatique de Paris 6, Paris, France

Abstract. On the one hand, Conceptual Graphs are widely used in natural language processing systems. On the other hand, information fusion community lacks of tools and methods for knowledge representation. Using natural language processing techniques for information fusion is a new field of interest in the fusion community. Our aim is to take the advantage of both communities and propose a framework for high-level information fusion. Conceptual Graphs model contains aggregation operators such as join and maximal join. This paper is dedicated to the extension of the maximal join operator in order to manage heterogeneous information fusion. Domain knowledge has to be injected into the maximal join operation in order to satisfy the constraints of fusion. The extension relies on relaxing the equality constraint on observations and on using fusion strategies. A case study illustrates our proposition and we describe the experimentations that we conducted in order to validate our approach.

1 Introduction

The first step of the decision-making process is to get information in order to elaborate a decision from it. Such a process is difficult as information is distributed across various sources and on different media. A lot of studies concern the fusion of either low-level data or data expressed through the same media. Our aim is to concentrate on high-level and heterogeneous information fusion. Even if some papers report about how to use ontologies to store domain knowledge ([1]), the Information Fusion community lacks techniques able to model knowledge. The objectives of our work is thus to propose an approach and a framework dedicated to high-level and heterogeneous information fusion. By high-level information, we mean that our aim is to manipulate semantic objects.

Conceptual graphs [2] are a widely used formalism for knowledge representation. The advantages of using graph structures, and particularly conceptual graphs model, to represent information have been stated in [3]. The authors explain how criminal intelligence information and model can effectively be stored as conceptual graphs. We propose to take advantage of this representation and go further by using the same model for information fusion. Using the same model for both information representation and information fusion has a major advantage. It allows us to remove the bias due to the translation from one formalism to another when using distinct models.

Among all the operators that were defined on the conceptual graphs structures, we are particularly interested in the maximal join. Maximal join allows the fusion of two graphs that are not strictly identical. We propose to use it in order to fuse different descriptions of a single object of the real world. Maximal Join must nevertheless be extended. Domain knowledge is widely used in the information fusion community in order to solve conflicts during fusion. Therefore, we propose to introduce some domain knowledge inside the maximal join operation.

Section 2 presents related works as well as the case study that we used to illustrate our proposition. The use of the conceptual graphs formalism for fusion is described in section 3. In particular, we detail in this section the suitability of maximal join operator for high-level information fusion. Section 4 details our proposition of extension for the maximal join. This extension relies on the use of external fusion strategies detailed in the same section. We describe in section 5 the experimentation that we conducted on the case study, in order to validate our approach. We then conclude and present future work.

2 Context

2.1 Related Work

Our aim is to use the output of intelligent sensors as input observations for our system. For textual information, these intelligent sensors are systems able to analyze the meaning of the texts and store it as machine readable information. As conceptual graphs were initially developed in order to analyze natural language, a lot of studies exist ([4], [5], [6]), aiming at transforming textual information items into conceptual graphs. Considering other media, studies such as [7] and [8] have been realized. They aim at automatically analyzing images and videos and store the resulting descriptions as conceptual graphs. Finally, as stated in [9] and [10] conceptual graphs are widely used to formalize several domains of knowledge as different as biomedical risks or corporate modeling. Therefore, we use conceptual graphs for knowledge representation. Furthermore, we propose to go beyond the usual use of conceptual graphs and take advantage of conceptual graphs operators for information fusion.

The information fusion community is more involved in studies aiming at fusing low level data. The use of techniques and methods taken from natural language processing is a new field of interest in the fusion community (see [11] and [12] for instance). People look at how to use ontologies to model a domain. We claim that conceptual graphs are a good candidate for information fusion since the formalism contains the maximal join operator and the structures are easily understandable.

2.2 Case Study

The approach that we propose can be applied to any domain for which a model can be drawn *a priori* and stored as an ontology. In order to validate it on real

data, we used a real world case study that concerns TV program descriptions. The purpose is to fuse descriptions given by different sources. Our aim is to obtain more complete and precise descriptions of the TV programs and to get a better scheduling of the programs.

Our first source of information (called DVB stream) is the live stream of metadata associated with the video stream on the TNT (Télevision Numérique Terrestre). The DVB stream gives descriptions of TV programs containing schedule and title information. It is very precise about the begin and end times of programs and delivers information about the technical characteristics of the audio and video streams.

The second source of information is an online TV magazine. The descriptions contain information about the scheduling of the programs, their titles and the channels on which they are scheduled. They also contain more details about the contents (summary of the program, category, list of actors and presenters etc).

3 Using Conceptual Graphs for Information Fusion

Conceptual Graphs [2] is a formalism particularly well suited to represent knowledge in a media- and source- independent way. We briefly introduce the way we will use it for information fusion.

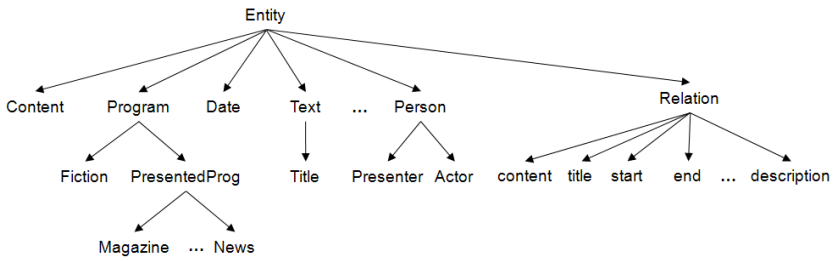

Fig. 1. Type hierarchy for TV programs

Defining the domain model is the first step of the fusion process. First, the ontology of the domain is defined. Figure 1 depicts a subset of the type hierarchy that was defined for the TV program case study. Then, the set of situations that are expected to happen are formulated through the canonical basis. Potential interactions between the entities (defined as concepts and relations in the ontology) are represented using conceptual graph structures. Figure 2 shows an example of an abstract canonical graph. It describes the model of a TV program.

After defining the domain model, we automatically acquire the observations into the conceptual graph formalism. Figure 3 show example of observations that were made on DVB stream and telepoche.fr website and stored as conceptual graphs.

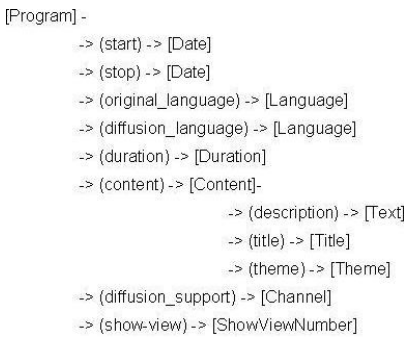

Fig. 2. TV Program Model

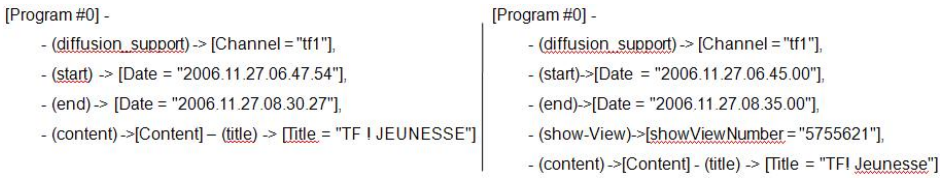

Fig. 3. Observations on DVB stream and telepoche.fr

Maximal Join is a major function in the process of fusion of conceptual graph structures. Two compatible sets of concepts from two different conceptual graphs are merge into a single one. There may be several possibilities of fusion between two observations, according to which combinations of observed items are fused or not. This phenomenon is well managed by the maximal join operator, as joining two graphs maximally results in a set of graphs, each one of it being a fusion hypothesis.

4 Towards a Framework for Information Fusion

4.1 Extending Maximal Join operator

Maximal join is a fusion operator which has to be modified in order to manage observations coming from different sensors. These observations may depict different points of view or different levels of detail and abstraction. The values of the concepts may be different while representing several observations of the same object.

Figure 4 gives an example of such a case. The maximal join of the two graphs G1 and G2 results in G3. The two concepts [Date: "2006.11.27.06.45.00"] and [Date: "2006.11.27.06.47.54"] cannot be joined using the standard maximal join operator as their values are different. However, because we know the domain that is modeled here, we have clues to say that the two concepts still represent the same entity in the real world. A TV program has only one begin time and there

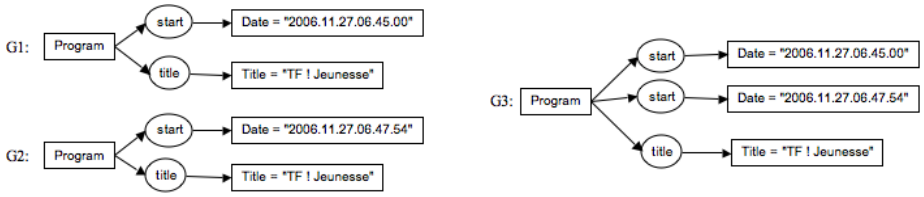

Fig. 4. Limitation of maximal join

are often slight differences between the times given by different sources. Fusion heuristics must be added in the maximal join operation. Therefore, the notion of compatibility between concepts is extended from compatible conceptual types to compatible referents and individual values. The domain knowledge necessary to this extension is stored as compatibility rules that are called Fusion Strategies.

4.2 Fusion Strategies

As explained before, the notion of compatibility between concepts in the maximal join operation has to be extended in order to support information fusion. Real data is noisy and knowledge about the domain is often needed in order to fuse two different but compatible values into a single one. Therefore, we introduced the notion of fusion strategies. They are rules encoding domain knowledge and fusion heuristics. We use them to compute the fused value of two different observations of the same object. On the one hand, the fusion strategies extend the notion of compatibility that is used in the maximal join operation. According to some fusion strategy, two entities with two different values may be compatible and thus fusable. On the other hand, the strategies encompass functions that give the result of the fusion of two compatible values.

Fusion strategies integrating domain knowledge and operator's preferences are the intelligent part of our fusion system. These strategies are implemented as $IF < conditions > THEN < fused - value >$ rules. They take conceptual graphs and conditions on the concepts as premises. The conclusion is a conceptual graph that integrates functions defining the values and referents of its concepts.

5 Validation

We implemented a fusion platform based on the approach that we propose. The platform was developed in JAVA and uses the AMINE platform ([13]) as a service provider for conceptual graphs definitions and basic manipulations. The fusion strategies are rules that were implemented as independent JAVA classes.

5.1 Experimentation

As detailed before, the domain that we chose in order to validate our proposition concerns TV program descriptions. The aim is to obtain as much TV program

descriptions as possible, concerning the TV programs scheduled on a TV channel, during one day. Furthermore, these descriptions should be as precise as possible with regards to the programs that were effectively played on the channel.

In order to compare the result of the fusion to the programs that were really performed, we collected TV program descriptions from the INAthèque. The INA, Institut National de l'Audiovisuel ([14]), collects the descriptions of all the programs that have been broadcasted on the French TV and radio. The exact begin and end times of the different programs are recorded. First, we know whether a fused program corresponds to the program that was really played. Second, we compare the times that were processed by fusion to the real diffusion times.

During one day, we request every 5 minutes the two sources of information to give us the next scheduled program on one channel. The two provided TV program descriptions are then fused using one of the fusion strategies. Once the fusion is done, we make sure that the description follows the general model for TV program descriptions. For instance, if the program has two different titles, it means that the fusion failed and the resulting description is rejected.

The well formed descriptions are then compared to the reference data. If they are compatible, the fused program description is considered to be correctly found with regards to reality. If the description is either badly formed or any part of the description doesn't correspond to the reference data, we consider that the program wasn't correctly found. For correctly found programs descriptions, we then compare the computed begin and end times to the real ones.

We measured the quality of the fusion that we obtained using different strategies. Therefore, we launched our experimentations using the fusion platform first combined with no strategy and then with three different ones. The first experiment -no fusion strategy- is equivalent to using the maximal join operator for information fusion. The three fusion strategies are the following:

Strategy 1 extends dates compatibility. Two dates are compatible if the difference between the two is less than five minutes. If two dates are compatible but different, the fused date should be the earliest one if it is a "begin date" and the latest one otherwise.

Strategy 2 extends dates and titles compatibility. The dates compatibility is the same as the one of strategy 1. Two titles are compatible if one of them is contained in the other one, after removing typography clues (upper cases, punctution marks...).

Strategy 3 extends dates and titles compatibility. The dates compatibility is the same as the one of strategy 1. Two titles are compatible if the total length of common substrings between the two exceeds a given length, after removing typography clues.

5.2 Results

We present here the results that we obtained during our experimentation. We first looked at the percentage of programs that were correctly found, according

	TF1	France4	France5	BFM	Gulli	iTV	M6	NRJ12	NT1
% prog found - no strategy	0	15,52	59,7	7,15	1,13	37,67	0	0	0,4
% prog found - strategy 1	0,18	18,53	73,47	7,15	1,13	42,71	0	0	9,66
% prog found - strategy 2	51,49	92,24	90,34	42,29	80,23	73,97	32,5	36,16	77,05
% prog found - strategy 3	64,23	92,24	80,46	30,08	71,35	82,99	47,71	37,5	81,88

Fig. 5. Percentage of programs correctly fused and identified with different strategies

to the different strategies that we used. Figure 5 shows the results we obtained on a representative selection of TV channels.

As expected, we can see that the fusion of observations using the maximal join operation only is not sufficient. Only the descriptions with strictly identical values are fused. There is too much noise in real data for a fusion process that doesn't take into account some knowledge about the domain. Therefore, the three previously cited fusion strategies were applied. The more the compatibility constraints between two values are relaxed, the better the results are. This is obvious as it is equivalent to inject more and more knowledge about the domain and knowledge about the general behavior of objects in the external world.

A second interpretation of our results consisted in the observation of the time lag between the fused description and the reference ones. Figures 6 and 7 give examples of the results obtained on two different channels. Each point represents a program and is located in the grid according to the difference between the fused begin and end times and the real broadcasted times. On Figure 6 only three points are visible. Actually, only two programs were badly guessed and all the others are represented by the point with coordinates (0,0). On Figure 7 we can see that almost all the programs are starting after the fused begin time. This seems to be due to the fact that advertisement is scheduled at the beginning of the time slots dedicated to each TV program.

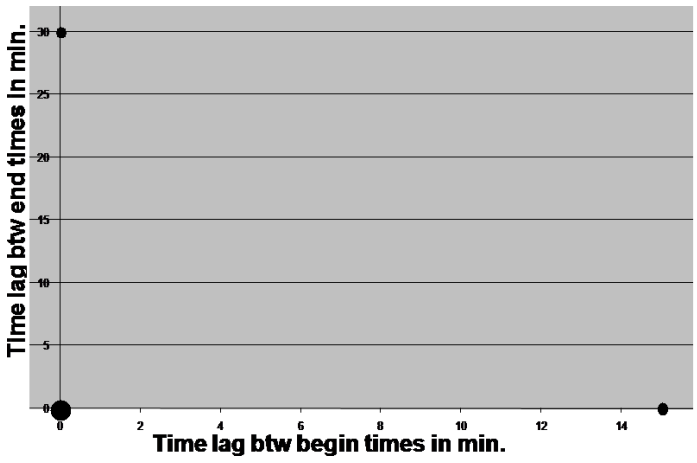

Fig. 6. Time lag between fused and broadcasted time on France 4 channel

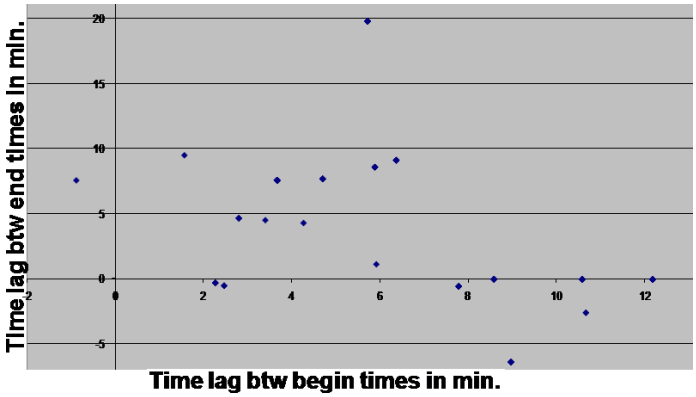

Fig. 7. Time lag between fused and broadcasted time on TF1 channel

The different experimentations that we carried out showed that the quality the fusion process is very heterogeneous, according to several parameters. First of all, it depends on the channel on which the observations are done. Some channels broadcast the programs almost always at the scheduled time, so the observations on both sources are identical and coherent with reality. In the meantime, most channels don't follow this rule. Then, the time of the day when the observation is made is important as well, as the specificity of the channel. For non popular channels and at times of low audience, we observed a lot of errors in the programs given by the TV magazine.

6 Conclusion

This paper proposes to use the conceptual graphs model for information representation and fusion. Using the same model for both purposes avoids the bias due to the translation from one formalism to another one. We detailed the extension that we proposed for the maximal join operator. This extension allows to fuse not strictly identical observations. It is based on the use of domain knowledge to relax the constraints when aggregating concepts. The standard maximal join is only based on structures and types compatibility. The extended version introduces the notion of fusion strategy. Fusion strategies are rules that allow to add a domain dependent notion to the fusion process. A case study was developed in order to illustrate and validate our approach on real data.

The first results of our study are promising as we showed that the use of the maximal join operation is relevant for information fusion. The operator must nevertheless be enriched with domain knowledge in order to be usful on real data which are noisy.

Current and future work will first deal with the study and improvement of the fusion strategies. In particular, we will focus on the use of the reliability of the information sources. Then, we will develop strategies that take the context of observation into account.

Finally, our approach can be used in other application domains. We are currently using the approach and the fusion platform on a crisis management case study concerning the Ivory Coast events. Information items are extracted from newspaper articles and then fused in order to obtain a global representation of the situation in the country at different dates.

References

1. C. Matheus, M. Kokar and K. Baclawski. A Core Ontology for Situation Awareness. 6th International Conference on Information Fusion, Cairns, Queensland, Australia, 2003, pp. 545-552.
2. J. F. Sowa, Conceptual Structures. Information Processing in Mind and Machine, Addison-Wesley, Reading, MA, 1984
3. R. N.Reed, and P. Kocura, Conceptual Graphs based Criminal Intelligence Analysis, in Contributions to 13th International Conference on Conceptual Structures, 2005, pp. 146-149
4. P. Zweigenbaum, and J. Bouaud., Construction d'une représentation sémantique en Graphes Conceptuels partir d'une analyse LFG, 4ème Conférence sur le Traitement Automatique des Langues Naturelles, Grenoble, France, 1997, pp. 30-39.
5. J. Villaneau, J-Y. Antoine, and O. Ridoux, LOGUS : un système formel de compréhension du franais parlé spontané-présentation et évaluation, 9ème Conférence sur le Traitement Automatique des Langues Naturelles, Nancy, France, 2002, pp. 165-174.
6. M. Montes-y-Gomez, A. Gelbukh, A. Lopez-Lopez, Text mining at detail level using conceptual graphs, 10th international conference on conceptual structures, Borovets, Bulgaria, 2002, pp. 122-136.
7. P. Mulhem, and W. K. Leow, and Y. K. Lee, Fuzzy Conceptual Graphs for Matching Images of Natural Scenes, 7th International Joint Conference on Artificial Intelligence, Seattle, Washington, USA, 2001, pp. 13971404.
8. M. Charhad, Modèle de Documents Vidéo basés sur le Formalisme des Graphes Conceptuels pour l'Indexation et la Recherche par le Contenu Sémantique, Thèse de L'universit J. Fournier, Grenoble, 2005.
9. F. Volot, and M. Joubert, and M. Fieschi, Knowledge and Data Representationwith conceptual graphs for Biomedical Information Processing : a Review, Methods Inf Med., N37 pp. 86-96, 1998.
10. O. Gerbe, and B. Guay, and M. Perron, Using Conceptual Graphs for Methods Metamodeling, 4th International Conference on Conceptual Structures, Bondi Beach, Sydney, Australia, 1996, pp. 161-175.
11. F. Deloule, D. Beauchêne, P. Lambert, B. Ionescu, Data Fusion for the Management of Multimedia Documents, 10th international Conference on Information Fusion, Quebec, Canada, 2007.
12. M. Gagnon, Ontology-based Integration of Data Sources, 10th international Conference on Information Fusion, Quebec, Canada, 2007.
13. AMINE Platform: http://amine-platform.sourceforge.net/
14. INAthèque: http://www.ina.fr/archives-tele-radio/universitaires/index.html