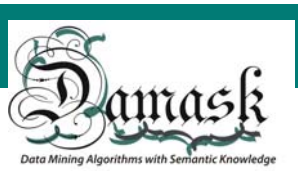

TIN2009-11005
DAMASK

Data-Mining Algorithms with Semantic Knowledge

PROYECTO DE INVESTIGACIÓN
PROGRAMA NACIONAL DE INVESTIGACIÓN FUNDAMENTAL,
PLAN NACIONAL DE I+D+i 2008-2011
ÁREA TEMÁTICA DE GESTIÓN: Tecnologías informáticas

Deliverable D1.2

Data Types

Authored by
Montserrat Batet, Universitat Rovira i Virgili
David Isern, Universtitat Rovira i Virgili
Aïda Valls, Universitat Rovira i Virgili

Document information

project name:	DAMASK	
Project reference:	TIN2009-11005	
type of document:	Internal report to be included in further deliberables	
file name:		
version:		
authored by:	M. Batet, D. Isern, A. Valls	15/04/2010
co-authored by		
released by:		. .200
approved by:	Co-ordinator	Antonio Moreno

Document history

version	date	reason of modification
1.0	08.March.2010	Priminary hierarchy of data types differ from the listed ones on the DAMASK proposal
	13.April.2010	Detailed description of the different data types
	. .200	
	. .200	
	. .200	
	. .200	
	. .200	
	. .200	

Table of Contents

1 Introduction

The first task of DAMASK called T1 - *Semantic integration of the information available in heterogeneous Web resources* includes two preliminary subtasks, tasks 1.1 and 1.2 (see Figure 1). The former task discusses the related work in information extraction distinguishing algorithms to extract structured, semi-structured and non structured resources. The latter is focused on analyzing which are the different types of data that can be extracted using the previous algorithms.

In this document we present which are the types of data that are considered in the literature, concerning recommendation, data mining and decision making processes. For each of them, a definition is done and some examples in the tourism domain are given.

As it will be shown in this report, currently there is available a large set of types of data (*e.g.*, mainly numerical and linguistic). In DAMASK project only a subset of them will be considered. This subset has been selected taking into account that the information should be obtained in an automatic and unsupervised way form quite complex resources available in the Web. This limits the possibility of obtaining some types of information. After presenting the list of data types that will be handled in the project, the document provides an example of their usage.

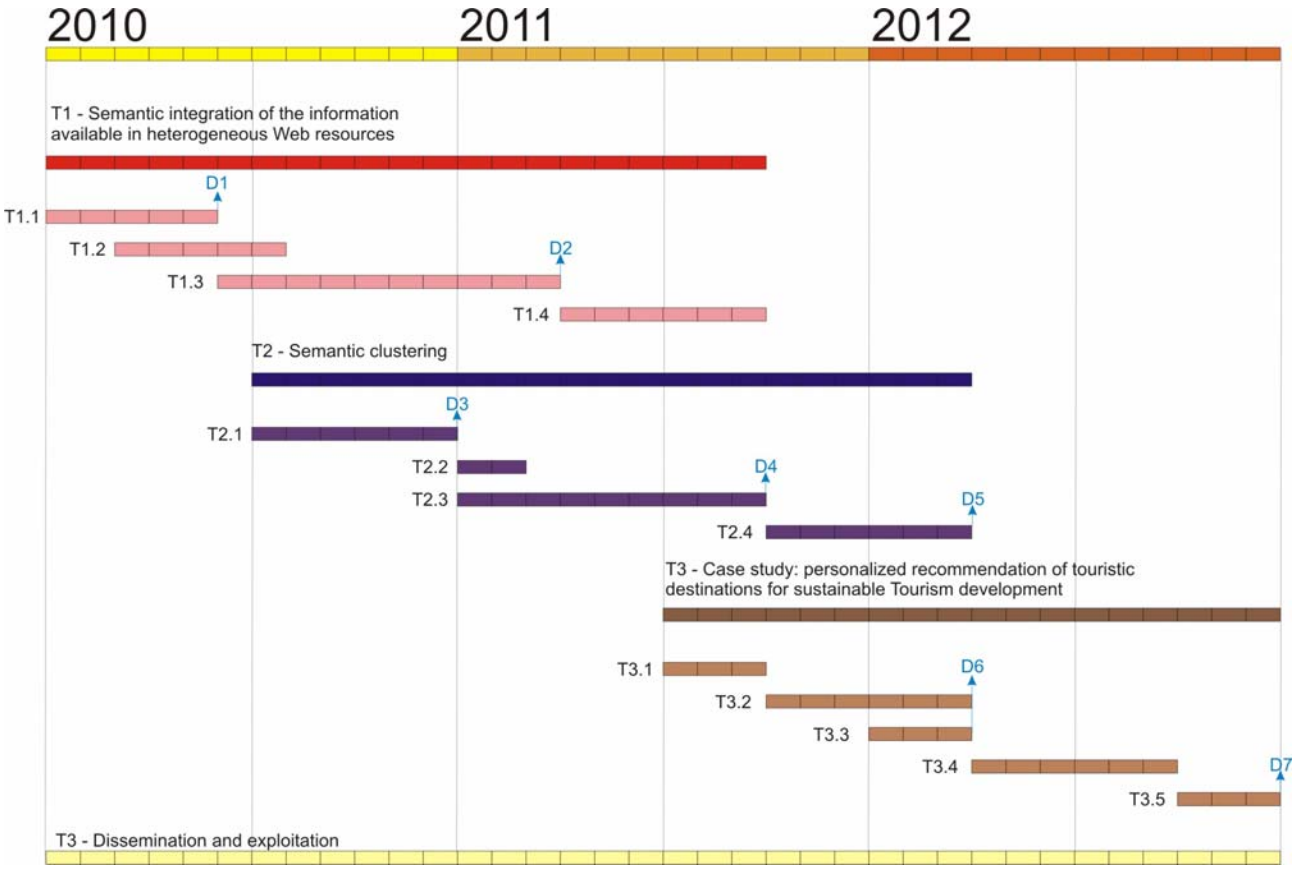

Figure 1: Tasks of DAMASK

2 Survey of data types

Although in data mining and decision making the types of values that are considered are usually the same, the nomenclature and organization changes from one author to another. (*e.g.*, (Brouwer, 2008), (Leacock & Chodorow, 1998), and (Torra & Narukawa, 2007)). Due to that there is not a common nomenclature and classification, in this document we provide a compendium of all the data types, proposing a particular denomination for each one, which will be adopted in the DAMASK project. Moreover, we present a classification according to the ones given in the literature.

The most traditional classification divides the data into two main groups: quantitative and qualitative. The former are also known as *numerical*, and the second is also known as *categorical*. However, with the growth of the Web and other digital sources of information, other types of data are being used. For example, we can have *multimedia* resources (photos, videos, audio) or different kind of *textual* data.

In this project we will not consider multimedia resources. The rest of data types are organized into the following 5 different typologies:

1) ***Numerical***: values that belong to some predefined subset of numbers. The main characteristic of this type of variable is that the values can be managed easily using arithmetic operators and have a direct comparison operator defined on them.

This type of data can be refined into four types:

a) *Measurement*. Indicates a quantity. It usually corresponds with a Real number (including Natural, Integer, Positive and Negative). For instance, the number of inhabitants of a city will be given using a Positive value.

b) *Ordinal*. The numerical values do not indicate a quantity but the position of the element in a ranking. They define a scale that permits to rank the objects, where the difference between adjacent values on the scale is not necessarily equal. For example, suppose that a group of tourists was asked to make a ranking of 5 events from the best to the worst. The event at position 1 is the better than the second and that the third, but we cannot know what is the strength of the difference (i.e. preference) between the events at position 1 and 2, or between the events at positions 2 and 3. In this case, arithmetic operations cannot be used. We can use operators based on comparisons of elements or some basic statistical measures such as percentiles and medians.

c) *Ratio*. *Ratio* scales indicate a proportion between two quantities. They are commonly encountered in physics. In this case, the measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. All statistical and mathematical numerical operations can be applied. In addition, they can be used to in measures that require a ratio, such as studentized range or coefficient of variation.

d) *Interval*. This permits to give an imprecise numerical value to one variable. The imprecision is expressed by indicating an interval of continuous values. For example, we can have that the average

temperature in Tarragona is [15-20] Celsius degrees. To handle this kind of values, the usual mathematic operations should be extended to deal with intervals (Rodgers & Nicewander, 1988).

With the *interval* scale we come to a form that is "quantitative" in the ordinary sense of the word. Almost all the usual statistical measures are applicable here (*e.g.* mean, standard-deviation, rank-order correlation, product-moment correlation), unless they are the kinds that imply a knowledge of a 'true' zero point.

e) *Fuzzy numbers*. Another way to manage the uncertainty is by using Fuzzy Sets theory (Zadeh, 1965). In this case, a membership function is associated to each number in a given subset of the Naturals. This permits to work with approximate numbers. For example, if the quality of the beaches in some place is 8, it means that it is “around” 8 as it is displayed in Figure 2. The fuzzy sets theory has a wide range of operators that permit to work with this kind of values (*e.g.* T-norms, T-conorms, T-indistinguishabilities, etc.).

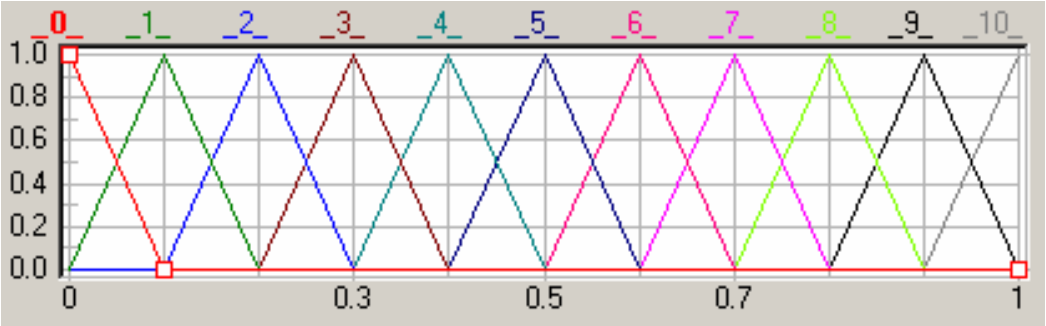

Fig 2. Fuzzy numbers

2) ***Booleans:*** is a primitive data type having one of two values: true or false, intended to represent the truth values of logic and Boolean algebra, or to represent the presence or absence of some feature. Matching coefficients are used to measure the similarity between to Boolean (*i.e.*, binary) variables (Sokal & Sneath, 1973). These measures are defined on the basis of the four values of the contingency table for the two Boolean variables (see Fig 3), such as the *Simple matching*, *Jaccard* or *Dice* coefficients (Dunn & Everitt, 1982). From these, the most popular one, in multiple disciplines, is still the *Jaccard* coefficient (Baets, *et al.*, 2009). These type of similarity measures were well defined and characterised by (Anderberg, 1973) and (Gower, 1971).

3) ***Categorical*** features take a set of values expressed with labels in a predefined set of terms (i.e. vocabulary). Some subtypes of categorical variables can be distinguished:

a) *Nominal*. A set of data is *nominal* if the values or observations belonging to it can be partitioned into a set of categories (i.e. modalities), which are mutually exclusive. For example, a destination has the characteristic of “language” with categories “English,” “French,” and so forth. Nominal features do not have any relationship between the values, except for equality among them. Hence the dissimilarity between two values is defined in terms of equality.

b) *Ordinal*. A set if data is *ordinal* if the values belong to categories that are totally ordered. For instance, cities can be classified according to their number of inhabitants in "village", "town", "city", "metropolis", "megalopolis". Operators are based on equality and ordering.

c) *Linguistic variables*: composed by *fuzzy sets*. In this type of elements, membership functions of the fuzzy terms distributed around a mid (*i.e.*, neutral) term. For example, when evaluating the users' satisfaction on some topic, linguistic labels like "very low", "low", "almost low", "medium", "high", "very high", "perfect" can be used. Aggregation and comparison operators have been defined to work with this type of data (Xu, 2008).

4) ***Semantic variables***. These variables have a non fixed and large set of possible values, without any order or scale of measurement defined between terms. These variable can be semantically interpreted with the help of additional knowledge, such as a domain ontology (Studer, *et al.*, 1998).

These variables can not be considered as *Categorical variables* as categorical variables have a predefined domain of terms (*i.e.*, modalities). The comparison between two values in categorical variables is simply based on the equality/inequality of words (and sometimes related with some kind of ordering of the categories), due to the lack of proper methods for representing the meaning of the terms.

Using *Semantic variables*, it is possible to establish different degrees of similarity between values (*e.g.*, in hobbies, trekking is more similar to jogging than to dancing) from a semantic point of view. These variables could be multi-evaluated. In fact, the computation of the semantic similarity between concepts is an active trend in computational linguistics (Jiang & Conrath, 1997; Leacock & Chodorow, 1998; Resnik, 1995; Wu & Palmer, 1994).

5) ***Texts***. When extracting information from documents, one can identify some paragraphs or short sections that seem to give a relevant description related to some particular feature (*e.g.*, the creation of the city). Those variables will have a short text associated. In order to be able of using this information in the clustering process, some textual analysis must be done to extract the relevant features of the text. The result will be a list of noun phrases that not necessarily can be directly associated to some particular feature. For example, consider the following text for the city Tarragona.

> "In Roman times, the city was named **Tarraco** (Ταρρακών) and was capital of the province of Hispania Tarraconensis (after being capital of Hispania Citerior in the Republican era). The Roman colony founded at Tarraco had the full name of **Colonia Iulia Urbs Triumphalis Tarraco**.
>
> The city may have begun as an Iberic town called *Kesse* or *Kosse*, named for the Iberic tribe of the region, the Cosetans, though the identification of Tarragona with Kesse is not certain. Smith suggests that the city was probably founded by the Phoenicians, who called it '*Tarchon*, which, according to Samuel Bochart, means a citadel. This name was probably derived from its situation on a high rock, between 700 and 800 feet above the sea; whence we find it characterised as *arce potens Tarraco*."

A vector of relevant words could be {Roman, capital, colony, Iberic tribe, Phoenicians, citadel, high rock}. Notice that the words must be analysed in an integrated way in order to be correctly compared with the

description of another city (Patwardhan & Pedersen, 2006). Moreover, some of these words could be identified with concepts of an ontology and be treated as semantic variables.

6) ***Others***. In (Hernández Orallo, *et al.*, 2004), there are listed other data types such as multimedia, temporal or spatial series, which are out of the scope of the DAMASK project, thus, they are not analysed in this survey.

As a final remark, it is important to note that an object that includes a set of criteria can contain *missing* values. This type of values will be represented using the symbol "?" into the data matrix.

3 Data types covered in DAMASK

In the previous section, a set of available data types have been introduced. In the DAMASK project we will only manage a subset of them that are showed in Table 1. We have selected a subset of data types that cover the different families of attributes, in particular: measurement, nominal and semantic types. As it has been explained, they have quite different characteristics, so they will involve the definition of different methods for the acquisition of those values from Web resources, and the management of quite different methods for similarity measurement.

TABLE 1
DATA TYPES USED TO MODEL DATA RETRIEVED FROM WEB RESOURCES

Attribute	Coverage
Numerical	
Measurement	Yes
Ordinal	No
Ratio	No
Interval	No
Fuzzy numbers	No
Boolean	No
Categorical	
Nominal	Yes
Ordinal	No
Linguistic variables	No
Semantic	Yes
Text	No
Others	No

As the information provided in the resources available in the Web is mainly in a textual format (structured or non-structured), some of the types of data not commonly found. In particular, ordinal data (both numerical and categorical) and values with uncertainty, like intervals, fuzzy numbers or linguistic variables. The consideration of ratios, texts and other types of data is left for future projects.

As an example, table 2 displays a matrix of objects as the ones that will be used in this project. It takes into account the data types that will be studied.

TABLE 2
EXAMPLE OF MATRIX OF OBJECTS AND CRITERIA

City	*Number of inhabitants*	*Continent*	*Climate*
	(Measurement)	(Nominal)	(Semantic)
Barcelona	4185000	Europa	Mediterranean
London	7556900	Europa	Temperate marine
New York	18800000	America	Humid continental
Reus	107118	Europa	Mediterranean
Tarragona	140223	Europa	Mediterranean
Paris	11769433	Europa	Oceanic

4 References

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York: Academic Press, Inc.

Baets, B. D., Janssens, S., & Meyer, H. D. (2009). On the transitivity of a parametric family of cardinality-based similarity measures. *International Journal of Approximate Reasoning, 50*, 104–116.

Brouwer, R. K. (2008). Clustering feature vectors with mixed numerical and categorical attributes. *International Journal of Computational Intelligence Systems, 1*(4), 285-298.

Dunn, G., & Everitt, B. (1982). *An Introduction to Mathematical Taxonomy*: Cambridge University Press.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*(4), 857-872.

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos*: Editorial Pearson.

Jiang, J., & Conrath, D. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*of the Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X)* (pp. 19-33). Taiwan.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification *WordNet: An electronic lexical database* (pp. 265-283): MIT Press.

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts*of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together* (pp. 1-8). Trento, Italy.

Resnik, P. (1995). Using Information Content to Evalutate Semantic Similarity in a Taxonomy. In C. S. Mellish (Ed.), *Proc. of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995* (pp. 448-453). Montreal, Quebec, Canada.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician, 42*(1), 59-66

Sokal, R., & Sneath, P. (1973). *Numerical Taxonomy*. San Francisco, US: WH Freeman.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *Data and Knowledge Engineering, 25*(1-2), 161-197.

Torra, V., & Narukawa, Y. (2007). *Modeling Decisions: Information Fusion and Aggregation Operators*: Springer Berlin Heidelberg.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection*of the Proceedings of the 32nd annual Meeting of the Association for Computational Linguistics* (pp. 133 -138). Las Cruces, New Mexico.

Xu, Z. (2008). Linguistic Aggregation Operators: An Overview In H. Bustince, F. Herrera & J. Montero (Eds.), *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models - Intelligent Systems from Decision Making to Data Mining, Web Intelligence and Computer Vision* (Vol. 220, pp. 163-182): Springer-Verlag Berlin Heidelberg.

Zadeh, L. (1965). Fuzzy sets. *Information and Control, 8*, 338–353.