User Focused Database Summarization Approach

Amel TRIKI
RIADI-GDL Laboratory
amel.triki@riadi.rnu.tn

Yann POLLET
CEDRIC-CNAM Laboratory
yann.pollet@cnam.fr

Mohamed BEN AHMED
RIADI-GDL Laboratory
mohamed.benahmed@riadi.rnu.tn

Abstract

Mining information from very large databases poses numerous challenges. In fact, systems that can mine such voluminous databases are increasingly desirable. In this context, we propose a generic approach of database summarization that takes into account the user's interest topic.

Innovation in our work consists in the generation of a set of database summaries having different levels of granularity in order to satisfy the user's expectations.

1. Introduction

The growth of databases volumes draw attention in the research field. In fact, for large volumes of information, manual analyses are no longer possible.

In this work, we propose a new approach to succinctly summarize very large databases.

Our proposal consists in building a set of database summaries that gives many levels of granularity in order to satisfy the user's requirements.

Our approach will be used in a medical context to summarize a medical knowledge database.

The remainder of this paper is structured as follows. The second section aims at introducing the database summarization field. The third section is devoted to the elicitation of the theoritical foundation of our proposal. The fourth section details the database summarization process from a pragmatic point of view. In the fifth section, we present an illustrative example of our approach through a case study related to a medical database. The last section exposes our future work and main conclusions.

2. Overview of the database summarization field

This section is devoted to the exploration of the database summarization field by defining as a first step what we mean by the term database summary; then, we present the different possible types of database summarization before investigating the different summarization techniques and the related work. As a fourth step, we point out the issues raised by database summarization and conclude the section with the main characteristics of a database summary.

2.1. Database summary definition

A database summary can be defined as a reductive transformation of the original database to a more concise form, through content reduction by selection and/or generalization of what is important in the source database.

So, the major goal of a summary is to give the main ideas of the original database but with a more concise form.

2.2. Different types of database summarization

As we defined earlier, summarization involves the selective elimination of the "unneeded" data. This last term is ambiguous since it is linked to the business objective of the database summary.

We can, therefore, identify mainly two types of summaries: topic oriented/ user focused / query focused summaries and generic summaries [1].

Topic oriented summaries concentrate on the user's desired topic of interest whereas generic summaries give a reduced representation of the original database with a similar coverage and a minimum of redundancies.

2.3. Different database summarization techniques and related work

A database summary can be obtained by extraction or abstraction.

Extraction techniques attempt to identify and retain the most relevant tupples in a database. In other words, extraction consists in the choice of some parts of the database at the expense of the rest. Extraction methods consist in selection and projection techniques.

Projection is a vertical reduction that eliminates some attributes while selection is a horizontal reduction that removes some tupples from the database [2] [3].
Although these two extraction techniques are profitable since they reduce rapidly database volumes, they present a major disadvantage; they don't take into account information gradation.

Database summrization can be also fullfilled by abstraction which is the process of generating new tupples that don't exist in the original database but that give a general and concise idea of the existing tupples by replacing them with related concepts. Summaries abstracts are usually generated after a preliminary extraction process in which the most representative tupples are recognized.

It is possible to distinguish mainly four categories of absraction techniques: Aggregation methods, Fuzzy set techniques, Attribute Oriented Generalization and Schema Summarization.

Aggregation methods consist in the process of gathering information and expressing it in a summary form, it can be fullfilled through techniques like On Line Analytical Processing (OLAP) and Data Cube or merely "GROUP BY" aggregate [4]. However, these techniques are more suitable for numerical attributes.

Fuzzy set techniques aim at replacing database tupples with fuzzy linguistic labels. In this field, we can mainly mention the researches led around the data summarization model SaintEtiq [5, 6, 7, 8, 9, 10, 11].

Attribute Oriented Generalization is based on the principle of replacing specific attribute value with more general concept user-defined beforehand [12].

Schema summarization is a technique that summarizes a database by the simplification of the database model structure [13].

The succinct overview of the database summarization topic raises many challenges and several problematic issues which are detailed in the following subsection.

2.4. Overview of the database summarization issues

Database summarization raises many issues which can be recapitulated as follows:

1) The focused database summary depends on the user's goal, thus, it is possible to generate several summaries.
2) A summary database is inevitably oriented to serve specific applications; hence, the summary depends on the context and purpose factors.
3) A database can change in the time by inserting, modifying and deleting tupples and tables. So, the interestingness of tupples can change which would affect the summary.
4) A summary can have many levels of granularity; it should contain enough information to satisfy the user's need and, at the same time, should not contain any redundant or superfluous information. So, how to find the good balance.
5) How to capture the important content and to identify cue tupples.
6) Some measures of interestingness to evaluate database summarization quality vary whether summaries are extracts or abstracts. In other words, measures appropriate to extract summaries could not be applicable to abstract summaries.

These observations reveal that it is essential to define clearly the characteristics of a database; these points are tackled in the next section.

2.5. Different database summarization characteristics

A database summary can be characterized by several descriptors which can be evaluated with different measures of interestingness. These measures can be classified as either objective or subjective [14][15].

Objective measures are based on the general structure of the summary whereas subjective measures are based on the user's expectations.

We present, in table 1, the main characteristics of a summary [14].

Table 1. Database summary characteristics and the corresponding evaluation measures

Database summary characteristic	Explanation
Complexity	Number of database components in terms of tables, relations and tupples.
Representativity	Ability of a summary to give a fair idea of the original database by covering the important content
Stability	Ability of a database to be represented by the most reduced extension
Robustness	Indicates whether noisy data can affect or no the generated summary.
Context sensitivity/ Coherence	Ability of a method to take into account contextual information

Once the database summarization field explored, it is henceforth possible to present the theoretical foundation of our approach as described in the next section.

3. Theoritical foundation of database summarization

In this section, we present the database summarization from a theoretical point of view. We present, firstly, the basic assumptions adopted in our work, secondly, we detail the database summaries representation.

3.1. Basic assumptions

As our scope is to deal with different databases which can be relational, object oriented, temporal,… we assume, in our work, that a database is given by a UML class diagram.

A database B can be represented by a set of classes {C1, C2, …, Ci…, Cn}, n, i ∈ **N** (natural numbers), Ci ∈ *C* (all the possible classes)

Each class is defined as a set of attributes {A1, A2,…, Ai,…, An} n, i ∈ **N** and Ai ∈ ***A*** (class attributes) where each attribute has a domain *DA*.

We notice that *DA* designates all the values that can be taken by the attributes which can be either quantitative or qualitative.

3.2. Representation of the database summaries: Embedded lattices

A database can generate many summaries with different levels of granularity.

The generated database summaries can be organized within a lattice of classes's properties *(E_FC,* $\subseteq$) in which each concept is itself the top of a lattice of the attributes's domain values. We define the two stages of lattice in the next subsections.

3.2.1. Definition of the classes's properties lattice.

We consider the formal context FCat = (O, A, I) which consists of:

Objects O : all the database instances (tupples).

Attributes A: all the database attributes.

I: binary relation $I \subseteq < O \times A >$: for each couple (o,a) of $O \times A$, *I* means that an object o is described by an attribute a.

For X $\subseteq$ O and Y $\subseteq$ A, we define:

X'={a ∈ A| a I o for all o ∈ X}

Y'={o ∈ O| a I o for all a ∈ Y}

(X,Y) is a formal concept of (O,A,I) iff

X $\subseteq$ O and Y $\subseteq$ A : X' = Y and X = Y'

So, X is the extent and Y is the intent of (X,Y).

We designate E_FC as the set of formal concepts of the context (O, A, I). (E_FC, $\subseteq$) defines a concept lattice. In fact, the inclusion $\subseteq$ is a partial order (reflexive, antisymmetric and transitive).

Every pair of concepts in this partial order has a unique greatest lower bound (meet). The greatest lower bound of (o_i, a_i) and (o_j, a_j) is the concept with objects $o_i \cap o_j$; it has as its attributes the union of a_i, a_j, and any additional attributes held by all objects in $o_i \cap o_j$.

Symmetrically, every pair of concepts in this partial order has a unique least upper bound (join). The least upper bound of (o_i, a_i) and (o_j, a_j) is the concept with attributes $a_i \cap a_j$; it has as its objects the union of o_i, o_j, and any additional objects that have all attributes in $a_i \cap a_j$. These meet and join operations satisfy the axioms defining a lattice.

3.2.2. Definition of the properties's domain values lattice.

Every concept C of the lattice *(E_FC, $\subseteq$),* described above, can be also represented by a lattice of the properties's domain values. We consider the formal context FC = (Op, Ap, Ip) in which:

Op designates the tupples describing the properties of a concept C of *E_FC.*

Ap designates the parts of the properties's domain values.

Ip is a binary relation $Ip \subseteq < Op \times Ap >$ which denotes that for every couple (o,a) of $Op \times Ap$, Ip means that o is described by the domain values of a.

Similarly to (*E_FC*, $\subseteq$), we can prove that the concepts of the formal context CF : E_FCp can be partially ordered by inclusion ($\subseteq$) and defines with the join and union opertors a lattice.

Once we have defined the database summaries organization from a theoritical point of view, we detail the process to generate summaries from a pragmatic point of view.

4. Database summarization generation process

In this section, we detail the different steps to generate a database summary that satisfies the user's expectations.

Thus, we define an important notion in our proposal, which is the summarization trace, then, we give an overview of the summarization generation process.

4.1. Summarization trace

As it is difficult to achieve, in "one shot", the focused summary with the desired granularity, we propose to generate summaries incrementally.

So, we define the summarization trace as the different iterations necessary to fulfill the desired summary.

In other words, the trace is the browse of the summaries lattice till the user is satisfied.

This notion of trace is crucial since it can be later used to accelerate the find of a summary with a certain granularity. In fact, for a similar user's request, instead of building a summary in multiple iterations, it is possible to reuse a trace.

The generation process is described in the following subsection.

4.2. Principle of the summary generation process

A database summary has to fulfill the user's requirements in terms of content and granularity. So, our proposal has to take into account these two aspects.

The content of the summary in term of classes is defined according to the request formulated by the claimant of information that precise classes having a certain interest from his point of view.

The granularity of the summary is defined by browsing the ontology /hierarchy /lattice… associated to every selected class.

In other terms a summary is a sort of "Cartesian Product" of the classes concerned by the database summary.

Stages permitting to get a summary define the database trace introduced earlier.

Our approach can be applied in multiple fields. In our case, we kept a special interest to the medical domain as explained in the next section.

5. Case study: Application of the summary approach on a medical database

5.1. Case study scope

Our research falls within a European project aiming at creating, managing and accessing medical knowledge for different user's profiles; medical experts, medical practitioners, medical students and patients.

In this project, a voluminous knowledge database would be handled; it includes information about diseases, medical guidelines, protocols, medicines, treatments, anonymous medical cases…

Two major issues are encountered: Firstly, mining information in such huge database is too difficult. Secondly, the information expected varies from a user's profile to another one, and even for the same profile, user's expectations can be very different.

Thus, it is important to provide a framework that allows database summarization according to user's focus.

Before introducing the framework, we present an overview of the medical database related to our case study.

5.2. Overview of an extract of the medical database to consider

We present an extract of the medical database described in the case study by the class diagram in figure 1.

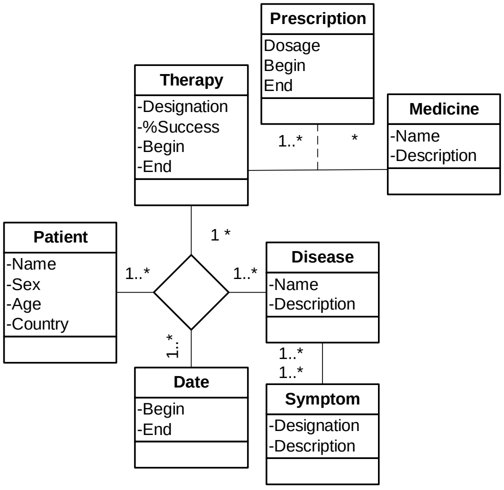

Figure 2. Extract of the medical database to summarize

Each class can be represented as follows:

Disease: Diseases are described through an ontology like SNOMED or BIOMED

Therapy: Therapies structured hierarchically, are extracted from anonymous patients.

Medicine: Medicines are described by Galen ontology.

Patient: Information on patients is taken from anonymous subjects.

Our approach will be supported by a framework as described in the next paragraph.

5.3. Overview of the database summarization framework

The database summarization framework has to satisfy some requirements namely:

1) The environment has to be graphical to ease the access to different and multiple users.
2) The framework has to integrate many tools for loading the database, visualizing it and generating the summaries.

3) Users have to manage a panel of choices to refine or generalize data using as well aggregation as fuzzy labels or other techniques.
4) The summarization trace must be saved to be reused later.
5) After formulating a request, the traces that respond to similar demands must be available.

Taking into account the mentioned points, we propose to develop an open Java platform that integrates a collection of tools to support the database summarization process from loading the database to generating the summaries. To provide the component of lattice visualization and manipulations, we will integrate the Galicia Software intended to construct interactively galois lattices.

6. Conclusion and future work

In this paper, we introduced our approach that aims at generating a user focused database summary.

Our proposal consists in building incrementally a summary by selecting key classes and attributes and refining or generalizing some properties values. This process leads to the construction of a summarization trace which is a collection of summaries that vary in granularity and content.

The different summaries that can be generated are structured within embedded lattices, so, the different iteration to look for a summary with a certain granularity can be viewed as navigation through the embedded lattices.

Our approach will be applied in a medical context and supported by a framework. The framework proposed, based on Galicia software, will integrate a panel of tools to load, visualize, manipulate and generate database summaries.

As future work, we aim at conceiving and developing the framework to put in practice our approach.

7. References

[1] D.R. Radev, E. Hovy, K. McKeown "Introduction to the special issue on summarization," *Computational Linguistics*, vol. 28, Issue 4, 2002, pp. 399–448.

[2] J.F. Roddick and al. "Methods and Interpretation of Database Summarisation". *DEXA*, 1999, pp. 604-615

[3] A. Lubinski. "A small database answers for small mobile resources", *International Conference on Intelligent Interactive Assistance and Mobile Multimedia Computing,* Rostock, November, 2000.

[4] J.Gray and al. "Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total". *Computing Research Repository (CoRR).* 1999.

[5] D. Lee, M. Kim. "Database summarization using fuzzy ISA hierarchies". *IEEE Transactions on Systems*, Man, and Cybernetics, Part B 27(1), pp. 68-78. 1997.

[6] D. Lee, M. Kim. "Discovering Database Summaries through Refinements of Fuzzy Hypotheses." *ICDE*. Pp.223-230. 1994.

[7] K. Janusz. "Fuzzy logic for linguistic summarization of databases". *Proceedings of the 8th International Conference on Fuzzy Systems (FUZZ-IEEE'99), Vol. 1 (pp. 813–818). Systems (IPMU'2000*), Vol. 2, pp. 1035–1040. 2000.

[8] G. Raschia, N. Mouadib. "SEQ: a fuzzy set-based approach to database summarization". *Fuzzy Sets and Systems*. Vo 129(2): pp. 137-162 2002.

[9] R. Saint Paul, G. Raschia, N. Mouadib. . "General Purpose Database Summarization". *VLDB*. Pp. 733-744. 2005.

[10] R. Saint Paul, G. Raschia, N. Mouadib. "Database Summarization: The SaintEtiQ System". *ICDE*. Pp. 1475-1476. 2007.

[11] D. Dubois, H. Prade. "Fuzzy sets in data summaries—outline of a new approach". *Proceedings 8th Int. Conf. on Information Processing and Managament of Uncertainty in Knowledge-based.* (2000).

[12] R.J. Hilderman, H.J. Hamilton. "Data Mining in Large Databases Using Domain Generalization Graphs". *Journal of Intelligent Information Systems,* Vol 3, Issue 3, pp.95 – 234, Nov-Dec, 1999

[13] C. Yu, H.V. Jagadish. Schema Summarization. *VLDB*, pp.319-330, 2006.

[14] R.J. Hilderman, H.J. Hamilton. "Principles for Mining Summaries Using Objective Measures of Interestingness". *J*

[15] D. Lee, M. Kim. "A Hypothesis Refinement Method for Summary Discovery in Databases". *Conference on information and knowledge Management (CIKM),* Washington, D.C, USA.. pp. 274-282. 1993.