A Data Warehouse Model for Integrating Fuzzy Concepts in Meta Table Structures

Daniel Fasel
Information System Research Group
Department of Informatics
University of Fribourg
Boulevard de Prolles 90
1700 Fribourg, Switzerland
daniel.fasel@unifr.ch

Khurram Shahzad
Information Systems Laboratory
Department of Computer and System Science
Royal Institute of Technology (KTH)
Forum 100, SE 164 40, Kista
Stockholm, Sweden
mks@dsv.su.se

Abstract

In classical data warehouses (DWH), classification of values takes place in a sharp manner, because of this true values cannot be measured and smooth transition between classes does not occur. In this paper, a fuzzy data warehouse (FDWH) modeling approach, which allows integration of fuzzy concepts without affecting the core of a DWH is presented. This is accomplished through the addition of a meta-table structure, which enables integration of fuzzy concepts on dimensions and facts, while preserving the time-invariability of the DWH and allowing analysis of data both sharp and fuzzy. A comparison to existing approaches for integrating fuzzy concepts in DWH is presented. Guidelines for modeling the fuzzy meta-tables and a meta-model for the FDWH are also outlined in this paper. The use of the proposed approach is demonstrated by a retail company example. Finally, a comparison of fuzzy and classical data warehousing approaches is presented.

1 Introduction

Over the years data warehouse technology has been used for analysis and decision making in enterprises [5]. Typically, a data warehouse gives a set of numeric values (called facts) that are based on a set of input values in the form of dimensions [7]. The numeric values of a classical data warehouse can be difficult to understand for business users, or may be interpreted incorrectly. Therefore, for a more accurate interpretation of numeric values, business users require an interpretation in meaningful non-numeric terms. However, if the transition between terms is crisp, true values cannot be measured and smooth transition between classes cannot take place [3]. To address this problem, fuzzy concepts as linguistic variables can be integrated with a data warehouse for the interpretation of numeric values [15].

Different approaches for the integration of fuzzy concepts in data warehouses have been proposed. Delgado et. al. [1] introduce fuzzy dimensions in which a dimension attribute is composed of a fuzzy concept. T-norm and t-conorm functions are used in this approach to aggregate over the fuzzy dimensions. Schepperle et. al. [12] and Krishna et al. [6] propose a similar approach of fuzzy dimensions but use normalized fuzzy membership degrees for the aggregation. Depending on how the fuzzy concept is generated, it is possible to lose sustainable information if the membership degrees are normalized. However, if fuzzy concepts are not normalized in these approaches, inconsistent results on aggregation may be produced and can harm the concept of summarizability [9]. Perez et. al. [11] introduce fuzzy concepts in the highest granularity of dimensions. Aggregation is therefore only executed over sharp dimension attributes. Feng et. al. [4] present a framework for integrating fuzzy concepts into a data warehouse including a discussion about the standard data warehouse operations over fuzzy concepts. This framework distinguishes between three layers of data cubes where the first layer describes sharp data, the second layer describes fuzzy facts and the third layer describes fuzzy dimensions. This framework only provides limited combinations of fuzzy and crisp elements and does not provide any schematic integration of meta-tables for fuzzy concepts. Fuzzy concepts may not be time-invariant and therefore all these approaches need major remodeling of the data warehouse if the fuzzy concepts change over time.

This paper proposes a fuzzy data warehouse model, which allows integration of fuzzy concepts directly into the data warehouse model. By using this approach, the concept of summarizability is not affected in dimensions as the
fuzzy concepts are rolled out in a meta-table structure. The proposed approach is more flexible as it allows integrating and redefining fuzzy concepts without the need for redesigning the core of a data warehouse. By using the fuzzy data warehousing approach, it is possible to extract and analyze the data simultaneously in a classical sharp and in a fuzzy manner.

In section 2 of this paper, a retail company example is used to demonstrate the problems of a classical data warehouse. In section 3, the fuzzy data warehouse model is proposed. Guidelines for modeling the fuzzy data warehouse are presented in section 4. The fuzzy data warehouse model is applied to the retail company example in section 5. Finally, section 6 presents a comparison of classical and fuzzy data warehouses.

2 A Retail Company Example

In this section, an example of a retail company from Switzerland and a classical data warehouse schema for the company is presented. Following the schema, example queries are executed, their results are presented and limitations of the classical data warehouse and existing fuzzy approaches are outlined.

The Switzerland-based retail company sells electronic devices, computers and servers to its customers in different parts of the country. It has four main suppliers (Apple, RIM, SUN, Nintendo) also called producers. The company divides Switzerland into different regions (North, East, South and West). The company’s products are categorized as either consumer electronics or business electronics.

2.1 Classical Data Warehouse for the Retail Company

The Snowflake schema [8] of the retail company’s data warehouse (formally, classical data warehouse) is shown in figure 1. Revenue is the fact and time, product and client are the dimensions which are used for the revenue analysis. For business analysis, the retailer can query its data warehouse with the advantage of standard data warehouse operations like slicing, dicing, roll-up and drill down. For example, the sum of the revenue per client in the year 2005 can be generated by applying a slice operation. The results of this query are shown in figure 2. Other interesting performance indicators for the retailer include the overall revenue per product category (see figure 3) and the overall revenue per region (see figure 4).

Several difficulties appear in the classical data warehousing approach given in figure 1. Products can only be categorized into one category i.e. the product iPhone can be categorized in consumer electronics or business electronics.

Figure 5 provides the technical characteristics used for classifying mobile phones into consumer electronics or business electronics. By using these technical characteristics, iPhone and BlackBerry are placed in different classes, as shown in figure 6. BlackBerry is not considered a consumer electronic because it does not have MP3 player and gaming capabilities. However, BlackBerry includes an internet browser and removable storage, which are the qualities of a consumer electronic. Therefore, it is not accurate to exclude BlackBerry from the category consumer electronics. Similarly, a city can belong to more than one region at a time e.g. Luzern is located in the central part of Switzerland and it may belong to all four regions (north, south, east, and west).

The limitations of the classical data warehouse approaches are summarized as follows:

- Classifications are handled sharp and therefore do not allow a smooth transition between classes.
- Decision-making processes are often verbal processes and domain specific terms are used for documenting decisions. A classical data warehouse approach does
• Facts are of quantitative nature. A qualitative interpretation of facts and dimension attributes is not supported in a classical data warehouse approach.

2.2 Limitations of existing Fuzzy Approaches

Using the fuzzy data warehouse approaches described in section 1, the limitations of the classical data warehouse approach can be overcome. However, several other limitations result from these approaches. In order to use fuzzy classification in a dimension hierarchy as described by [1], [12] and [11], the degrees have to be normalized to 1 to fulfill the summarizability. By applying normalization in this case, categorization information is lost.

Another limitation is the use of fuzzy concepts as dimensional attributes, which means a dimension can only be queried fuzzily, and a comparison between fuzzy and sharp queries is not possible.

In time, the business environment of a company is likely to change. In order to remain consistent with the business environment the fuzzy concepts have to be adapted regularly. By using fuzzy concepts as dimensional attribute, dimension structures have to be remodeled every time a change occurs.

The limitations of the existing fuzzy data warehouse approaches are summarized as follows:

• In order to use fuzzy dimensional attributes, the fuzzy membership degree must be normalized.

• Users must choose either a sharp or a fuzzy dimension but it is not possible to use fuzzy and sharp dimensions simultaneously.

3 Modeling a Fuzzy Data Warehouse

The purpose of this section is to present and describe a new fuzzy data warehouse model. In section 3.1 and 3.2 a set of concepts is defined and these concepts are used for defining the fuzzy data warehouse model given in section 3.3.

3.1 Fundamental Concepts defined

In order to propose the fuzzy data warehouse this section provides some fundamental concepts. These concepts are then further used to build the fuzzy data warehouse model.

Definition 1 (Domain of Attribute) A set of possible values or the range of possible values that an attribute can have is called domain of an attribute or universe of discourse. Domain D of an attribute A is represented by D(A).

Definition 2 (Target Attribute) An attribute or a metric that is required to be classified fuzzily is called a target attribute (TA). Under fuzzy classification, instances of an attribute TA are classified over a set (S) that is represented by a linguistic variable. The linguistic variable consists of a set of non-numeric terms called linguistic terms.
Definition 3 (Class Membership Attribute) A class membership attribute (CMA) for a target attribute TA, represented by CMA(TA), is an attribute that has a set of linguistic terms to which the target attribute may belong. The values of CMA are a set of distinct fuzzy classes to which the target attribute may belong. In other words, for all possible values of a target attribute (domain of attribute) there is a corresponding CMA value. The values of CMA are the values of the set S.

Definition 4 (Membership Function) A membership function of a target attribute TA is a function that is used to calculate the membership degree of a target attribute TA to a class membership attribute CMA.

Definition 5 (Membership Degree Attribute) A membership degree attribute MDA of a target attribute TA, is an attribute that has a set of membership degrees of the target attribute TA. The value of a membership degree is calculated by a membership function (as defined in definition 4) and is represented by, \( \mu(TA) = MD \); where MD is the membership degree of TA.

A fuzzy set for online customers has been defined by Werro [14]. Based on this definition the fuzzy set in this context is defined as follows:

If \( D(TA) \) is a set, then the fuzzy set CMA in \( D(TA) \) is defined as

\[
CMA = \{(TA, \mu_{CMA}(TA))\}
\]

Where \( TA \in D(TA) \), \( \mu_{CMA} : CMA \in [0, 1] \) is the membership function of CMA and \( \mu_{CMA}(TA) \in [0, 1] \) is the membership degree of TA in CMA.

3.2 Fuzzy Meta-Tables

According to the fuzzy classification approach proposed in this paper, there are at least two tables that are used for classification. An attribute that must be handled fuzzily is extended with two meta-tables. The first meta-table contains a description of the fuzzy concept and the second meta-table contains membership degrees of each instance with regard to the membership classes. One of the main advantages of this is that the numeric values and fuzzy classification using linguistic variables are accessible to users, i.e. the data can be analyzed sharp and fuzzy. The two tables are defined as follows:

Definition 6 (Fuzzy Classification Table) A table that consists of fuzzy classes and their unique identifiers is called fuzzy classification table. It is a two-attribute table that consists of an identity attribute and a class membership attribute, where the identity attribute is a unique identifier of the table values. Formally,

\[
FCT(TA) = \{(Identifier, CMA(TA))\};
\]

Definition 7 (Fuzzy Membership Table) A table that stores the values representing the degree to which a value is related to a fuzzy class is called fuzzy membership table. It is a table with four attributes: the identity attribute of the table, the identifier of the target attribute TA, the identifier of the fuzzy classification table \( FCT(TA) \) and membership degree attribute MDA for TA. Formally,

\[
FMT(TA) = \{(Identifier, Identifier of TA, Identifier of FCT(TA), MDA(TA))\};
\]

3.3 Fuzzy Data Warehouse Model

The fuzzy data warehouse model is a combination of four types of tables (see definition 8). These are dimension tables, fact tables, fuzzy membership tables and fuzzy classification tables.

One of the main advantages of this model is that both sharp and fuzzy classification can be combined, because the fuzzy concepts are rolled out of dimensions or facts and stored in meta-tables.

Definition 8 A fuzzy DWH model is a set of tables and it is represented by FDWH.

\[
FDWH = \{Dim, Fact, FCT(SA), FMT(SA)\},
\]

Where,
\text{Dim} = \{\text{a set of attributes, level of attributes}\}
\text{Fact} = \{\text{a set of facts, a set of measures}\}
\text{SA} = \{TA_1, TA_2, \ldots, TA_n\}

where \( n \) is the number of attributes that are to be classified fuzzily.

It is notable that the set of target attributes is a subset of set of dimension and set of facts. Formally, 
\( \text{SA} \) is a subset of \( \text{Dim} \cup \text{Fact} \) (i.e. for all \( TA_i \in \text{Dim} \cup \text{Fact}; i \) is from 1 to \( n \)).

For each attribute \( TA_1, TA_2, \ldots, TA_n \)
\( FCT(TA_i) = \{\text{Identifier, CMA}(TA_i)\}; \) where \( i = 1 \) to \( n \).
\( FMT(TA_i) = \{\text{Identifier, Identifier of FCT, Identifier of TA}_i, MDA(TA_i)\}; \) where \( i = 1 \) to \( n \).

4 Guidelines for Modeling the Fuzzy Data Warehouse

In this section, a set of guidelines for designing fuzzy data warehouse models and the usage of these guidelines for developing a meta-model for the fuzzy data warehouse (in section 4.4) is presented.

4.1 Distinct Fuzzy Classes / Linguistic Terms

A set of linguistic terms (also called fuzzy classes) is used for the classification of instances of a target attribute. In the simplest case, the linguistic terms are distinct i.e. there is a single set of linguistic terms with no repetition between them. In this case, one instance of a target attribute belongs to only one fuzzy class at a time and the degree of relation is measured by a membership function. Formally,

\( TA – instance \ 1 \ : \ Fuzzy\ Classes \ 1 \)

Guideline 1 Add a fuzzy classification table \( FCT \) and a fuzzy membership table \( FMT \) for each target attribute \( TA \), as shown below.

Guideline 2 If an instance of a \( TA \) belongs to a fuzzy class but with different membership degrees, add a \( FCT \) and \( M \) number of \( FMTs \), as shown below, where \( M \) is the number of distinct membership degrees.

4.2 Different Membership Degrees for the same Linguistic Terms

An instance of a target attribute may belong to a linguistic term (fuzzy class) but may have different degrees with which they belong to a linguistic term. This is because of the reason that multiple business users have different interpretations of a single instance of a target attribute i.e. multiple membership functions are used for a target attribute. Formally,

\( Attribute – instance \ 1 \ : \ Fuzzy\ Classes \ (1) \) but with different membership degrees

Guideline 3 If an instance of a \( TA \) belongs to multiple fuzzy classes but with the same membership degree, add \( M \) number of \( FCTs \) and a \( FMT \), where \( M \) is the number of distinct linguistic terms.

Guideline 4 If an instance of a \( TA \) belongs to more than one fuzzy class with different membership degrees, add \( M \) number of \( FCTs \) and \( FMTs \), as shown below, where \( M \) is the number of distinct linguistic terms, and one \( FCT \) is related to one \( FMTs \) at the most.
4.4 Meta model for Fuzzy Data Warehouse

By using the guidelines given in section 4.1, 4.2 and 4.3, a meta-model for the fuzzy data warehouse is presented. A fuzzy data warehouse consists of fact tables that have measures and key attributes. The key attributes represent identifiers as primary keys or foreign keys in the table structure. A fact has at least two or more dimensions and dimensions are composed of key or non-key attributes. The first dimension is a dimension time which is always given in a data warehouse model. Non-key attributes describe informational attributes of a dimension as i.e. the address or the name of a client in the dimension attribute client. Each attribute in a dimension has a level of attributes describing the hierarchy of the dimension.

Both facts and dimension attributes can be classified in accordance with fuzzy classes. Derived from guideline 2 (section 4.2), an attribute can have multiple fuzzy membership tables and each membership table can have one or more membership attributes. Additionally, derived from guidelines 3 and 4 (section 4.3), a fuzzy membership table may have multiple fuzzy classifications. The meta model for the fuzzy data warehouse is shown in figure 7.

Figure 7. Meta model for fuzzy data warehouse

Using this fuzzy data warehouse meta model, the fuzzy concepts of dimensions and facts can be propagated over the dimension hierarchies as described in [2] and [3]. The fuzzy concepts can also be combined in operations and queries with sharp dimension attributes or sharp measures as exemplified in figure 13 (section 5).

5 A Fuzzy Data Warehouse Example

In order to overcome the difficulties described in section 2, the data warehouse model of the retail company is transformed into a fuzzy data warehouse model. The fuzzy data warehouse snowflake schema for the retail company is presented in figure 9. The dimension attributes (region and product categorization) in the dimensions client and product can be redefined in fuzzy meta-tables. Also, a fuzzy concept on the fact revenue is defined. The concept classifies the revenue into three linguistic classes {high, middle and low revenue}. The linguistic concept of this fuzzy classification is shown in figure 8. The figure shows a smooth transition between the linguistic terms.

Figure 8. Linguistic concept revenue

In order to classify clients, a fuzzy concept on the dimension attribute client is defined. The customers can now be classified into three categories {A, B and C customers} where for each customer category the retailer has different sales strategies. A classical classification would consider all customers, who are producing 80% of the revenue as A customers, while the other customers would be divided into groups B or C. The drawback of this approach is that customers who have very close revenue (e.g. 79% & 80% of the revenue) are classified in different classes and have completely different priorities for the sales strategy. If the customers are classified fuzzily, each customer has a membership degree to all three classes and the retailer can adapt its sales strategies more accurately to the customers’ true business value.

For the classification of the fact revenue a fuzzy classification table is created to store the linguistic terms and a fuzzy membership table is created to store the membership degrees. The membership degrees are calculated by using an appropriate membership function as described in section 3, definition 5.

The dimension attribute product classification of the classical data warehouse described in section 2.1 (figure 1) can now be redefined as a fuzzy concept. With a fuzzy concept, a product can be classified into several categories depending on the applied classification method (as i.e. described in figure 5) and membership function. The two graphics shown in figure 10 represent the differences between a sharp and a fuzzy classification of products. In the
Figure 9. Fuzzy data warehouse snowflake schema

graphic Sharp Product Classification (represented by figure 10-a) a product only belongs fully to one class. Therefore, the iPhone 16GB and the Black Berry 81xx are classified completely differently despite being similar devices used in similar environments as stated in section 2.1. The graphic Fuzzy Product Classification (represented by figure 10-b) shows a fuzzy classification. The iPhone 16GB and the Black Berry 81xx now belongs to a certain degree to both classes. The fuzzy concept clearly demonstrates the similarity of these products and provides more accurate values of the products.

The fuzzy concept product category can be extended with another variation of linguistic terms, as described in guideline 2 (section 4.2). The first fuzzy classification table classifies the products with the linguistic terms business electronics and consumer electronics. By using another fuzzy classification table, the products can also be classified with the linguistic terms leisure products and professional products. Both linguistic concepts use the same fuzzy membership table to classify the products.

For the dimension attribute client, the linguistic concept with the terms A customer, B customer and C customer is defined in the fuzzy classification table. In the retail company example, classification is done in two different ways: a classification of the overall revenue of the client and a classification on a manually-created criteria catalogue from the sales department, allowing two fuzzy membership tables to be defined. The membership degree attributes in the first fuzzy membership table are calculated by using a membership function derived from the criteria catalogue. In the second fuzzy membership table, the membership degree attributes are calculated by using a membership function that takes the overall revenue of the clients into consideration.

Similar to a classical data warehouse, operations like slicing, dicing, drill-down and roll-up can be used on the fuzzy data warehouse. Figure 11 shows the result of a simple slice operation to extract all the products and their membership degrees for the product category (a fuzzy concept).

Based on this slice operation, the roll-up operation to the dimension attribute producer can indicate the affiliation of a producer to the terms of product category as shown in figure 12. More complex operations in the fuzzy data warehouse can be executed combining fuzzy and non-fuzzy dimensions and facts. Figure 13 shows the result of a dice
operation querying the data warehouse for all the customers that have a high revenue (fuzzy concept revenue) in consumer electronics (fuzzy concept product category) living in the region west (fuzzy concept region) in the year 2003 (sharp dimension time) listed by month and name.

This example demonstrates how the use of the fuzzy data warehouse model improves the retail company’s ability to classify its data more accurately to the needs of the company. Additionally, the fuzzy concepts can be adapted easily when the contextual environment of business changes (i.e. the criteria catalogue of the classification of the customers is changed). The meta-table structure allows defining and redefining the fuzzy concepts without any need to remodel the entire data warehouse structure.

6 Evaluation and discussion

In order to further demonstrate the use of the fuzzy data warehouse, a discussion on queries on a fuzzy data warehouse is presented. A comparison of classical and fuzzy data warehouses is also discussed.

6.1 Queries in a fuzzy data warehouse

For improving queries in the data warehouse an extended version of SQL can be used. fCQL [13] extends standard SQL in order to improve querying fuzzy concepts. In contrast to SQL, fCQL uses the linguistic context of the fuzzy concepts without the need to specify numerical values in the statement. Therefore, fCQL is a more human-oriented query language. fCQL has the following syntax [14]:

CLASSIFY Attribute list (similar to the select clause in SQL)
FROM Relation Name (similar to the from clause in SQL)
WHERE Selection Condition (similar to the where clause in SQL)
WITH Classification Condition (to integrate a fuzzy concept affiliation)
ALPHA Alpha-cut Condition (to modify the degree of the affiliation)

Figure 14-a presents a SQL statement for extracting all the low revenues of the year 2003 and figure 14-b presents the corresponding fCQL statement. In the with clause of the fCQL statement the fuzzy concept is included without using numeric values. To achieve the same constraint in the SQL statement, a where clause selecting the fuzzy concept and a second where clause giving the membership degree as a numeric value must be defined.

6.2 Comparison of Classical and Fuzzy Data Warehouse

The following table presents a comparison of the classical and the fuzzy data warehouse model in order to sum-
marize the main advantages of the fuzzy data warehouse approach.

<table>
<thead>
<tr>
<th>Classical Data Warehouse</th>
<th>Fuzzy Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>In a classical DWH, an instance does not belong to more than one class at a time. Because of this, true values of the classification can not be measured.</td>
<td>Classification of dimension attributes or facts in the FDWH is done in a fuzzy manner, allowing values to belong to more than one class and the classification to be more accurate.</td>
</tr>
<tr>
<td>Qualitative interpretation of facts and dimension attributes is not supported in a classical DWH.</td>
<td>A FDWH enables using non-numerical attributes. As a result, both qualitative and quantitative attributes can be used for analysis.</td>
</tr>
<tr>
<td>Decision-making processes are often verbal. A classical DWH approach does not include any linguistic concept to interpret the data.</td>
<td>The definition of linguistic variables can be derived from the business environment manually. This reduces the effort of interpreting numeric values and facilitates decision-making processes.</td>
</tr>
<tr>
<td>Only crisp data is used for analysis and decision making.</td>
<td>Both fuzzy and crisp data can be used for analysis and decision making.</td>
</tr>
<tr>
<td>The classical schema consists of dimensions and facts.</td>
<td>The FDWH schema consists of a classical schema together with fuzzy meta tables called fuzzy classification table and fuzzy membership table.</td>
</tr>
<tr>
<td>In a classical DWH only extracted data (slices, dices, etc.) can be classified. The classification can therefore not be propagated on other hierarchy levels of dimensions.</td>
<td>In FDWH the fuzzy concepts can be propagated over the dimensions in order to apply the classifications on other hierarchy levels.</td>
</tr>
<tr>
<td>The Retrieval of queries in a classical DWH is based on SQL in most cases.</td>
<td>A FDWH can be queried on a linguistic level. For example, FSQL (Meier et al. [10]) allows marketers to classify single customers or customer groups by classification predicates such as ‘loyalty is high and turnover is large’.</td>
</tr>
</tbody>
</table>

7 Conclusion

In this paper, a fuzzy data warehouse model that facilitates smooth transition between classes have been proposed. By using the fuzzy data warehouse model, data can be classified both fuzzily and sharply. Because of this, the FDWH supports qualitative and quantitative analyses, without affecting the core data warehouse schema. In addition, querying can be done based on natural language through direct use of the terminologies of the fuzzy classifications.

The study examines a retail company example for which a classical data warehouse is modeled, problems are defined and exemplified. A set of fuzzy data warehousing concepts are defined and explained. These concepts are then used to define the two types of tables used in fuzzy data warehousing, called fuzzy classification table and fuzzy membership table. The fuzzy classification table contains different classes according to which a classification takes place and the fuzzy membership table contains the membership degrees of the classified values. The concepts are furthermore used to define a meta model which gives a complete overview of the fuzzy data warehouse. The use of the fuzzy data warehouse has been demonstrated with the help of the retail company example, and by running sample queries. Finally, a comparison of a classical and a fuzzy data warehouse is presented.

Based on the study, the following can be concluded: a) the fuzzy data warehouse provides true values and a more accurate classification of values, b) by using fuzzy data warehouse, both qualitative and quantitative attributes can be used for analysis, c) both fuzzy and crisp measures can be used for analysis and decision making, d) Querying can include business environment terminology.

Future work should include real-time implementation of the proposed approach, evaluation of the success of the approach and qualitative measurement of its effect on enterprise analysis and decision making in complex systems.

References


