Soft Comput (2012) 16:135–151
DOI 10.1007/s00500-011-0740-1

ORIGINAL PAPER

A new model for linguistic summarization of heterogeneous data: an application to tourism web data sources

Ramón A. Carrasco · Pedro Villar

Published online: 8 June 2011

Abstract In this paper we present the problem of aggregating heterogeneous data from various websites with opinions about high end hotels into a database. We present the *fuzzy model based on the semantic translation* as a tool to obtain a linguistic summarization. The characteristics of this model (necessary to solve the problem) are not together on any of the existing linguistic models: the management of the input heterogeneous data (natural language included); the procurement of linguistic results with high precision and good interpretability; and the use of unbalanced linguistic term sets described by trapezoidal membership functions for defining the initial linguistic terms. We applied it to aggregate data from certain high end hotels websites and we show a case study using the high end hotels located in Granada (Spain) from such websites during a year. With this aggregated information, a data analyst can make several analyses with the benefit of easy linguistic interpretability and a high precision. The solution proposed here can be used to similar aggregation problems.

Keywords Data summarization · Fuzzy linguistic modelling · Opinion aggregation · Heterogeneous data integration

R. A. Carrasco (✉) · P. Villar
Department of Software Engineering,
University of Granada, 18071 Granada, Spain
e-mail: racg@ugr.es

P. Villar
e-mail: pvillarc@ugr.es

1 Introduction

With the rapid development of Web2.0 that emphasizes the participation of users, websites such as (Atrapalo 2011; Booking 2011; eDreams 2011; Expedia 2011; TripAdvisor 2011; Trivago 2011) encourage users to express opinions on tourism services like hotels by posting feature ratings and textual reviews (forums, news groups, etc.). These numerical ratings are often used by recommender systems to recommend highly rated hotels, assisting users in making decisions. The general approach (including the aforementioned websites) is to compute only the accurate numerical information given by users to provide a ranking value of these hotels and their features. However, a large portion of users do not provide feature ratings. Simply because it may cost users too much effort to provide detailed feature ratings. For example in TripAdvisor (2011) website, approximately 43% of users do not provide such ratings (Long et al. 2009). In this case, the opinions expressed by the users of tourism services in natural language form are an important source of information.

There is a method for aggregating these textual opinions called *Opinion Aggregation* (Hu and Liu 2004; Morinaga et al. 2002; Carenini et al. 2005; Ku et al. 2006; Miao et al. 2009; Zhang et al. 2009; Lazzari et al. 2009; Tang et al. 2009; Tsytsarau and Palpanas 2010; Ribeiro et al. 2002). The difference among opinion aggregation and other summarization tasks is the necessity to provide summaries along several features, aggregated over one or more dimensions. This problem imposes certain challenges related to the extraction of representative features and the calculation of the average sentiment or rating. The final goal though, is to determine the overall opinion of the community on some specific product, rather than the individual user opinions on the same product. The problems that have been studied in relation to opinion aggregation are mainly formulated around the aggregation of product reviews. Multiple architectures have been proposed to tackle this problem, the most popular ones (Hu and Liu 2004; Tsytsarau and Palpanas 2010) follows the steps outlined below (Fig. 1):

- *Collect*. Extracting information from the web data sources into a repository.
- *Identify*. This process starts with the identification of opinionated phrases, which may additionally involve a collection of phrase patterns. Identified phrases are then passed on to the feature extraction step, which may exploit a product taxonomy database (Carenini et al. 2005) to improve its results.
- *Classify*. Features and opinionative phrases are used in the sentiment classification step, which outputs sentiment polarities to aggregate over frequent features at the opinion aggregation step.
- *Aggregate*. In the last step, the opinions are aggregated per feature.

Our proposal is to aggregate feature ratings (step 4) by making use of users' textual reviews and numerical ratings (not only accurate data, but also approximate and interval values) from various websites with opinions about high end hotels. The main requirement of the problem is obtaining this aggregation with higher levels of accuracy while maintaining good linguistic interpretability.

Many aspects of different activities in the real world cannot be assessed in a quantitative form, but rather in a qualitative one, i.e., with vague or imprecise knowledge. In that case, a better approach may be to use linguistic assessments instead of numerical values. The fuzzy linguistic approach was introduced by Zadeh (1975). It is a tool used to model qualitative information in a problem. It is based on the concept of linguistic variable and has been used successfully in many problems (Bordogna and Passi 1993, 2001; Herrera-Viedma et al. 2007; Herrera-Viedma 2001; Delgado et al. 1992). Briefly speaking, linguistic variables are variables whose values are not numbers but words or sentences in a natural or artificial language. Linguistic variables have been used to linguistic data summarization (Yager 1982, 1991; Kacprzyk et al. 2000; Kacprzyk and Yager 2001; George and Srikanth 1996; Kacprzyk 1999; Kacprzyk and Zadrozny 2000; Laurent 2003) as well.

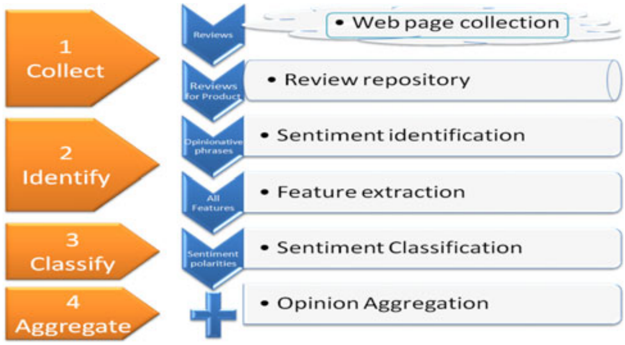


Fig. 1 Architecture of the opinion aggregation

Therefore, the fuzzy linguistic approach seems an appropriate framework for solving our problem (Lazzari et al. 2009). However, the necessary characteristics to solve this problem are not together on any of the existing linguistic models. These features are:

- *High precision and good interpretability of the results*. There is a limitation of most of linguistic models imposed by their information representation model and the computation methods used when fusion processes are performed on linguistic values. This limitation is the loss of information caused by the need to express the results in the initial expression domain that is discrete via an approximate process. This loss of information implies a lack of precision in the final results from the fusion of linguistic information. However, in our problem, the result of this aggregation should be expressed in a linguistic form with high precision. This characteristic is very important to make analyses of the integrated information. Thus, for example, a user could analyze the temporal evolution of a characteristic of a hotel, e.g., if this feature becomes more or less *good* for a period of time. This high precision should not be against the easy linguistic interpretability of the results.
- *The management of heterogeneous data*. Commonly included into the website pages to integrate: crisp (accurate data), approximate, intervals, several linguistic semantics, missing and undefined values, etc.
- *The management of unbalanced linguistic term sets described by trapezoidal membership functions*. Several authors consider that linear trapezoidal membership functions are good enough to capture the vagueness of the linguistic terms (Delgado et al. 1992). We can find an example of using this representation on a similar problem in Ribeiro et al. (2002). The management of unbalanced linguistic term sets is necessary because the experts want to express a higher resolution in the top of the linguistic scores (e.g. *good*, *excellent*, etc.) regarding to the bottom of the linguistic scores (e.g. *unacceptable*, *poor*, etc.). In the real problem considered, the high scores are very usual because we are dealing with high end hotels.

Our objective is to define a new fuzzy model that fulfils these characteristics, and then, applied it to a linguistic summarization in order to solve the proposed aggregation problem (aggregate phase in Fig. 1).

In this paper, for this purpose, we present the fuzzy model based on the semantic translation (FMST). Given an ordered set of unbalanced primary linguistic terms specified with trapezoidal membership functions, the basic idea consists to define a semantic translation of such terms and then obtain an ordered set which include the primary terms and semantic translations of such terms. After specifying this more precise representation model, we define the corresponding computation model including aggregation operations essential for application to the summarization model definition. If we are aggregating the age of hotel guests, the result could be, for example "*teenager* −2" with the linguistic interpretability "2 years to *teenager*". Therefore, FMST permits easy linguistic interpretability and a high precision to aggregate linguistic terms, owing to the semantic translation (−2 in the previous example). However, we need to extend it to manage more heterogeneous data (not only linguistic) included into the website analyzed, and then use this model to define a linguistic summary of data.

This new summarization model is used as aggregate phase in the opinion aggregation architecture specified in Fig. 1. We have used this entire architecture to aggregate (over a time dimension) heterogeneous data from various high end hotels websites. In particular, we show an example of the application using the high end hotels located in Granada (Spain) from such websites during the year 2009. Besides, we show examples of analyses that a data analyst can make using the easy linguistic interpretability and a high precision of the model.

The paper is structured as follows: Sect. 2 revises the preliminaries concepts, i.e., opinion aggregation, the fuzzy linguistic and data summarization approaches. Section 3 presents the new model, i.e., the FMST; we show as this model can manage more heterogeneous data (not only linguistic); and finally, we use this model to define a linguistic summary of data, in order to apply it to the problem to be solved. Section 4 presents a case study of the high end hotels aggregation. Finally, we point out some concluding remarks and future work.

2 Preliminaries

In this section we present the basic elements needed to understand our new proposal: opinion aggregation, fuzzy linguistic and data summarization approaches.

2.1 Related work on opinion aggregation

Most existing works in opinion aggregation follow the steps we listed in the Sect. 1" (see Fig. 1). This is the case of the method proposed by Hu and Liu (2004). They describe a system that aims at discovering words, phrases, and sentiments that best characterize some product. However, this pattern is not unique. For example, Morinaga et al. (2002) reversed the ordering of steps 1 and 2, and the experiments revealed that their system achieves a similar performance. Different approaches to feature extraction have been proposed. Hu and Liu (2004) identify features by building a list of noun-noun phrases using a natural language processing (NLP) parser, and then determining the most frequent ones. However, their approach outputs many irrelevant words and should be used in conjunction with other methods, as was suggested by Carenini et al. (2005). Accordingly, they introduce a domain taxonomy in the form of user-defined features, which are used to annotate data for training a feature classifier. Opinions are then collected and aggregated based on the full set of features, which consists of features extracted automatically (unsupervised learning) and also through the classifier (supervised learning). Alternatively, Ku et al. (2006) proposed a system that identifies features using information retrieval methods. They use a score per paragraph and per document, and a dictionary to determine polarity. The intuition here is that relevant features appear frequently in few of the paragraphs of many documents, or in many of the paragraphs of few documents. Aggregation of opinions has been traditionally performed over all the documents in some collection. Miao et al. (2009) used a time-decaying aggregation, retrieving only the most recent reviews that were marked by users as helpful. Zhang et al. (2009) introduced a novel approach, which interactively aggregates and displays sentiments based on different granularities of time and space (geographical location). Besides, the fuzzy linguistic approach has been used for solving this problem (Lazzari et al. 2009). This work also proposes this linguistic approach to solve the aggregation step of the most commonly used architecture (see Fig. 1). This entire architecture is too used in the case of study proposed in this paper. For an extensive survey of the area of opinion aggregation, the interested reader should refer to the works by Tang et al. (2009) and the Tsytsarau and Palpanas (2010).

2.2 The fuzzy linguistic approach

Since the concept was introduced (Zadeh 1975), linguistic variables have been widely used. Briefly speaking, *linguistic variables* are variables whose values are not numbers but words or sentences in a natural or artificial language; and these values of linguistic variables are called *linguistic labels*. In more specific terms, a linguistic variable is characterized by a quintuple $\langle H, T(H), U, G, M\rangle$ in which:

- H is the name of the variable.
- $T(H)$ is the term-set of H or the collection of linguistic values (labels).
- U is the universe of discourse.
- G is the syntactic rule, i.e., a context-free grammar which generates the terms in $T(H)$.
- M is the semantic rule which defines the meaning of each linguistic label X, $M(X)$, where $M(X)$ denotes a fuzzy subset of U.

The fuzzy linguistic approach (Zadeh 1975) is a tool used for modelling qualitative information in a problem. It is based on the concept of linguistic variable and has been satisfactorily used in many problems, such as, information retrieval (Bordogna and Passi 1993, 2001; Herrera-Viedma et al. 2007; Herrera-Viedma 2001), decision-making (Delgado et al. 1992; Yager 1999), etc. We have to choose the appropriate linguistic descriptors for the term set and their semantics. In order to accomplish this objective, an important aspect to analyze is the "granularity of uncertainty", i.e., the level of discrimination among different counts of uncertainty. Typical values of cardinality used in the linguistic models are odd ones, such as 7 or 9, where the mid term represents an assessment of "approximately 0.5", and with the rest of the terms being placed symmetrically around it (Bonissone and Decker 1986). Once the cardinality of the linguistic term set has been established, the linguistic terms and its semantics must be provided:

- *Generation of the linguistic terms*. Mainly, there are two possibilities to accomplish this task (Bordogna and Passi 1993; Bonissone 1982; Yager 1995). One of them involves directly supplying the term set by considering all the terms distributed on a scale on which a total order is defined (Herrera et al. 1995; Yager 1995). The other of them specify a context-free grammar G defined by the 4-tuple (Bordogna and Passi 1993): $\langle V_T, V_N, P, I\rangle$ where

  (a) V_T is the set of the terminal symbols, also called the alphabet.
  (b) V_N is the set of nonterminal symbols.
  (c) P is the set of the production rules.
  (d) I is the start symbol or axiom.

- *Semantic of the linguistic terms*. Often, the semantics of the terms are represented by fuzzy numbers, defined in the interval [0, 1], described by membership functions. A way to characterize a fuzzy number is to use a representation based on parameters of its membership function (Bonissone and Decker 1986). The linguistic assessments given by the users are just approximate ones. Some authors consider that linear trapezoidal membership functions are good enough to capture the vagueness of such linguistic assessments (Delgado et al. 1992). The parametric representation is achieved by the 4-tuple $[\alpha, \beta, \gamma, \delta]$ where β and γ indicate the interval in which the membership value is 1, with α and δ indicating the left and right limits of the definition domain of the trapezoidal membership function (Bonissone and Decker 1986). A particular case of this type of representation are the linguistic assessments whose membership functions are triangular, i.e., $\beta = \gamma$. Some authors (Bouchon-Meunier and Yao 1992) introduce a modifier that leads to a decrease or an expansive of the degrees of membership.

Next we analyze the models of fuzzy linguistic approach that we use in our system:

- *The approximative computational model based on the Extension Principle* (Bonissone and Decker 1986). This model uses fuzzy arithmetic based on the Extension Principle to make computations over the linguistic variables. This model can present the results in two ways: by means of the fuzzy numbers obtained from the fuzzy arithmetic computations based on the Extension Principle; or by means of linguistic labels computed from the fuzzy numbers obtained by performing a linguistic approximation process.
- *The ordinal linguistic computational model* (Herrera et al. 1996; Delgado et al. 1993). This symbolic model makes direct computations on labels, using the ordinal structure of the linguistic term sets $S = \{s_i\}$, $i \in \{0\ldots g\}$. Its results are inherently linguistic labels due to either the operators used, basically *max* and *min* operators or because in the computations on the order index there exist an approximation by means of the round operator.
- *The 2-tuple fuzzy linguistic approach* (Herrera and Martínez 2000) is a continuous model of representation of information that has been used in many applications (Moreno et al. 2010; Herrera-Viedma et al. 2007). The linguistic computational model based on linguistic 2-tuples carries out processes of ''computing with words'' without loss of information (typical of the other fuzzy linguistic approaches). It uses the 2-tuple fuzzy linguistic representation model and its characteristics to make linguistic computations, obtaining as results linguistic 2-tuples. A linguistic 2-tuple is defined by a pair of values (s_i, α_i), where $s_i \in S$ whose membership functions are assumed to be of the triangular type and $\alpha_i \in [-0.5, 0.5)$ represents the value of the symbolic translation. Roughly speaking, this symbolic translation supports the "difference of information" between a counting of information $b \in [0, g]$ obtained after a symbolic aggregation operation and the closest value in $\{0\ldots g\}$ that indicates the index of the closest linguistic term in S ($i = round(b)$).

2.3 Data summarization using fuzzy logic

The recent growth of Information Technology has implied, among others, the availability of a huge amount of data. Unfortunately, the availability of data does not make by itself the use of those data more useful and productive. Data summarization attempts to reduce facts to knowledge to aid decision making.

The linguistic summary can be viewed as a natural language like sentence that subsumes the very essence (from a certain point of view) of a set of data (Yager 1982, 1991; Kacprzyk et al. 2000; Kacprzyk and Yager 2001; George and Srikanth 1996; Kacprzyk 1999; Kacprzyk and Zadrozny 2000; Laurent 2003). This set is assumed to be numeric, usually large and not comprehensible in its original form by the human being. Often, the following context for linguistic summaries mining is assumed:

- $Y = \{y_1, \ldots, y_n\}$ is a set of objects (records) in a database, e.g., the set of hotel guests;
- $C = \{C_1, \ldots, C_r\}$ is a set of attributes characterizing objects from Y, e.g., place of residence, age, etc. in a database, $C_j(y_i)$ denotes a value of attribute C_j for object y_i, and $C_j(Y)$ denote the set $\{C_j(y_i)\}$ $\forall i \in \{1 \ldots n\}$.

Yager (1982, 1991) (Kacprzyk et al. 2000; Kacprzyk and Yager 2001) proposed that a linguistic summary of data set Y for a attribute C_j can be made in terms of three values (s_{Cj}, Q_{Cj}, T_{Cj}):

- A *summarizer* $s_{Cj} \in S_{Cj}$ i.e. an attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute C_j (e.g. "young" for attribute "age"). The set S_{Cj} contains all the possible linguistic terms defined for the attribute C_j.
- *A quantity in agreement* Q_{Cj}, is a proposed indication of the number of pieces of data that satisfy s_{Cj}, i.e. a linguistic quantifier (e.g. most).
- *Truth* (validity) T_{Cj} *of the summary*, i.e. a number from the interval [0, 1] assessing the truth (validity) of the summary (e.g. 0.7); usually, only summaries with a high value of T_{Cj} are interesting. Thus, the linguistic summary may be exemplified by "T_{Cj} (most of hotel guests are young) = 0.7".

To obtain the truth value T_{Cj}, the following procedure is used:

(1) $\forall i \in \{1 \ldots n\}$ calculate $s_{Cj}(C_j(y_i))$, the degree to which $C_j(y_i)$ satisfies the label s_{Cj}.

(2) *Calculation of the value* q_{Cj} *an indication* (*relative or absolute*) *of the number of pieces of data that satisfy such label* s_{Cj}. When Q_{Cj} is a relative quantity:

$$q_{Cj} = (1/n) \sum_{i=1}^{n} s_{Cj}(C_j(y_i))$$

and when Q_{Cj} is an absolute quantity:

$$q_{Cj} = \sum_{i=1}^{n} s_{Cj}(C_j(y_i))$$

(3) *Calculation of the truth value* T_{Cj} as the membership of q_{Cj} in the proposed quantity in agreement:

$$T_{Cj} = Q_{Cj}(q_{Cj})$$

3 Heterogeneous linguistic summarization using the new fuzzy model based on the semantic translation

Regarding the problem to be solved, the fuzzy linguistic models introduced in the Sect. 2.2 are considered constrained in several aspects:

- The approximative model based on the Extension Principle and the ordinal linguistic model, they both have a limitation: the loss of information caused by the need to express the results in the initial expression domain that is discrete via an approximate process. This loss of information implies a lack of precision in the final results from the fusion of linguistic information.
- The 2-tuple fuzzy linguistic approach eliminates the loss of information of the previous fuzzy linguistic approaches, owing to the symbolic translation. Therefore, this model has led to the objective of giving more accuracy to linguistic fuzzy modelling without losing interpretability. Besides, this model provides an easy mathematical formalism to deal with non-homogeneous information (Herrera et al. 2005). However, it represents only uniform and symmetrical distribution linguistic term set described by triangular membership functions.

3.1 The fuzzy representation model based on the semantic translation

Let $S = \{s_i\}$, $i \in \{0 \ldots g\}$ be a linguistic term set, such that each term s_i has associated the semantic of the trapezoidal membership function $[\alpha_i, \beta_i, \gamma_i, \delta_i]$. Let the fuzzy operator be σ: $S \times S \rightarrow [0, 1]$ such that $\forall s_i, s_j \in S$, $\sigma(s_i, s_j)$ represents *fuzzy degree of superiority* of s_i *over* s_j.

Definition 1 The operator σ forms a total order relation on S if fulfils:

$$\forall s_i, s_j \in S \quad \text{if } \sigma(s_i, s_j) > \sigma(s_j, s_i) \Leftrightarrow i > j.$$

There are many possible ways to define the operator σ over $A = [\alpha_A, \beta_A, \gamma_A, \delta_A]$ and $B = [\alpha_B, \beta_B, \gamma_B, \delta_B]$ (two trapezoidal possibility distributions, e.g. see Fig. 2): possibility and necessity theory (see Table 1), even the subjective criterion of some decision maker (Carrasco et al. 2001), etc.

A *triangular norm* (*t-norm* for short) (Klement et al. 2000; Petrík 2010) is a function t: [0, 1] × [0, 1] → [0, 1], such that $\forall x, y, z \in [0, 1]$ the following four axioms are satisfied: (i) *commutativity*: $t(x, y) = t(y, x)$, (ii) *associativity*: $t(x, t(y, z)) = t(t(x,y), z)$, (iii) *monotocity*: $t(x, y) \le t(x, z)$ whenever $y \le z$, (iv) *boundary condition*: $t(x, 1) = x$. Since Definition 1, we have semantic consistence on S regarding the previous order relation, i.e., we can say that S *fulfils a property of t-transitivity* for a t-norm t (Klement et al. 2000):

$$\forall s_i, s_j, s_k \in S, \sigma(s_i, s_k) \ge t(\sigma(s_i, s_j), \sigma(s_j, s_k)).$$

Definition 2 We define *the translation $d \in \Re$ of a term s_i* $\in S$ as following:

$$S_i + d = \begin{cases} [\alpha_i - (\delta_i - \gamma_i) + d, \beta_i + d, \gamma_i + d, \delta_i + d], & \text{if } i = 0, d > 0 \text{ and } \alpha_i = \beta_i \\ [\alpha_i + d, \beta_i + d, \gamma_i + d, \delta_i + (\beta_i - \alpha_i) + d], & \text{if } i = g, d < 0 \text{ and } \gamma_i = \delta_i \\ [\alpha_i + d, \beta_i + d, \gamma_i + d, \delta_i + d], & \text{otherwise} \end{cases}$$

We call *d as the value of translation* of s_i and it represents the "difference of information" between $s_i + d$ and s_i.

In the above definition, were considered as exceptional cases the existence of L and Γ type trapezoidal functions in the extreme left ($i = 0$) and right ($i = g$), respectively, i.e., the end result of the translation has been defined as symmetric (see Fig. 3).

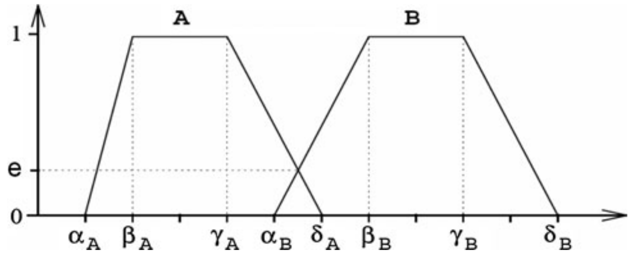


Fig. 2 Possibility operator definition $\varepsilon(A, B) = e$

We can conclude that a order relation defined over S is not fulfilled for all possible translation of the terms of S. In order to have semantic consistency on the possible translation of the terms of S we define:

Definition 3 We define the *maximum* (đ_i) and *minimum value* ($\underline{d}_i$) *of translation of a term $s_i \in S$, based on an operator σ and a threshold χ* as following:

$$\text{đ}_i = \begin{cases} 0, & \text{if } i = g \\ \text{Sup}\{d \in U/\sigma(s_{i+1}, s_i - 2 \times d) \\ > \chi\}, & \text{otherwise} \end{cases}$$

$$\underline{d}_i = \begin{cases} 0, & \text{if } i = 0 \\ \text{Sup}\{d \in U/\sigma(s_i - 2 \times d, s_{i-1}) \\ \ge \chi\}, & \text{otherwise} \end{cases}$$

Therefore, we can conclude that $\forall s_{j-1}, s_j, s_{j+1} \in S$, $j \in \{1 \ldots g-1\}$ *fulfil*:

$$\sigma(s_j - \underline{d}_j, s_{j-1} + \text{đ}_{j-1}) > \chi \text{ and } \sigma(s_{j+1} - \underline{d}_j, s_j + \text{đ}_{j-1}) > \chi.$$

Example 1 A set S of seven terms on the age of the hotel guest could be given as follows: s_0 = *baby*, s_1 = *child*, s_2 = *teenager*, s_3 = *young*, s_4 = *adult*, s_5 = *mature* and s_6 = old with the semantic of the unbalanced trapezoidal membership functions defined in the Fig. 4. Using the *possibly operator* shown in Table 1 for σ and the threshold χ equal to 0.75, it is obvious the superiority of the term adult on the term *young*, i.e., σ(*adult*, *young*) > 0.75. If we applied the maximum value of translation of the term *young* (đ_3) and minimum value of translation of the term *adult* ($\underline{d}_4$), the superiority of the translated term adult on the translated term *young* is still fulfilled, i.e., σ(*adult*$-\underline{d}_4$,*young* + đ_3) > 0.75 (see Fig. 5).

Definition 4 We define the set D_{STi} of the *semantic translations of the of a term $s_i \in S$* as following:

$$D_{\text{STi}} = \{d_i : d_i \in [-\underline{d}_i, \text{đ}_i]\}$$

Table 1 Examples of definition of the operator σ

Possibility operator definition σ		Necessity operator definition σ	
$= 1$	if $\gamma_A \ge \delta_B$	$= 1$	if $\alpha_A \ge \delta_B$
$= \frac{\delta_A - \gamma_B}{(\delta_B - \gamma_B) - (\gamma_A - \delta_A)}$	if $\gamma_A < \delta_B$ & $\delta_A > \gamma_B$	$= \frac{\beta_A - \gamma_B}{(\delta_B - \gamma_B) - (\alpha_A - \beta_A)}$	if $\alpha_A < \delta_B$ & $\beta_A > \gamma_B$
$= 0$	otherwise	$= 0$	otherwise

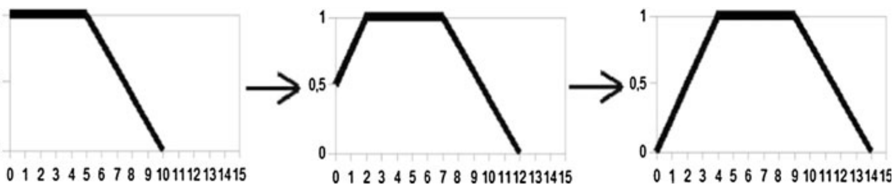


Fig. 3 Translation of a L type trapezoidal function for $d = 2$ and $d = 4$

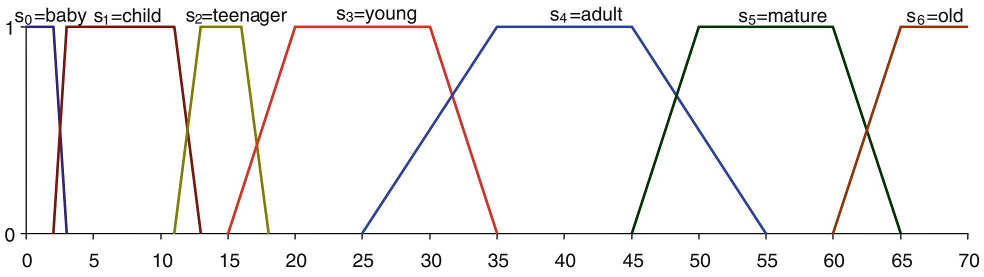


Fig. 4 Membership functions defined for set S

Let h be a natural number, in order to define a grammar that generates a finite set of terms we discretize the set D_{STi} according the following definitions:

Definition 5 We define the set D_{HSTi} of the *h-higher semantic translations of a term* $s_i \in S$ as following:

$$D_{\mathrm{HSTi}} = \{\text{đ}'_{ij} : \text{đ}'_{ij} = j \times \text{đ}_i / h\}, \quad \forall j \in \{1 \ldots h\}$$

We define the set D_{LSTi} of the *h-lower semantic translations of a term* $s_i \in S$ as following:

$$D_{\mathrm{LSTi}} = \{\underline{\text{d}}'_{ij} : \underline{\text{d}}'_{ij} = (-j + h + 1) \times \underline{\text{d}}_i / h\}, \quad \forall j \in \{1 \ldots h\}$$

Now we proceed to define the representation model as a set of the production rules defined in an extended Backus Naur Form in which the square brackets enclose optional elements, the symbol * indicates the possible repetition of the elements which follow, and the symbol | indicates alternative elements.

Definition 6 *The representation of the model based on the semantic translation* is generated from the context-free grammar G where, $\forall i \in \{0 \ldots g\}$ and $\forall j \in \{0 \ldots h-1\}$:

```
V_T =  s_i
V_N=   {<term>, <low comp term>, <hig comp term>, <trans low>,
       <trans hig>}
P =    <term> ::=  {<low comp term>|<primary term>|<hig comp term>}
       <low comp term> ::= *[<primary term> <trans low>]
       <trans low> ::= <sign trans low> {d̲'_ij|d̲'_ij|…| d̲'_ij}
       <hig comp term> ::= *[<primary term> <trans hig>]
       <trans hig> ::= <sign trans hig> {đ'_ij|đ'_ij|…|đ'_ij}
       <primary term> ::= s_i
       <sign trans low > ::= " - "
       <sign trans hig> ::= " + "
I =    <term>
```

Therefore, each <primary term> has associated the semantic of a trapezoidal membership function; and each <hig comp term> and <low comp term> have associated the semantic of the <primary term> with a difference of information of <trans hig> (higher) and <trans low> (lower), respectively.

3.2 The fuzzy computational model based on the semantic translation

The grammar G has led to the definition of a new ordinal set $\hat{S} = \{s_0, s_0 + \text{đ}'_{01}, \ldots, s_0 + \text{đ}'_{0h}, \ldots, s_g - \underline{\text{d}}'_{g1}, \ldots, s_g - \underline{\text{d}}'_{gh}, s_g\}$. Given that the primary terms have semantic translation 0, we will proceed to rename the set as $\hat{S} = \{s_0 + 0,\ s_0 + \text{đ}'_{01},\ \ldots,\ s_0 + \text{đ}'_{0h},\ \ldots,\ s_g - \underline{\text{d}}'_{g1},\ \ldots,\ s_g - \underline{\text{d}}'_{gh},\ s_g + 0\}$. Therefore, $\hat{S} = \{s_i + d_i\},\ i \in \{1 \ldots m\}$ and $m = (2 \times h + 1) \times (g-1)$; and if each term $s_i + d_i$ is renamed as $\hat{s_i}$ we have that $\hat{S} = \{\hat{s_i}\}, i \in \{1 \ldots m\}$. We can also conclude that the operator σ forms a total order relation on $\hat{S}$. We will define the computer model on this new set. This model will be more accurate as the number of semantic translations (h) is greater. Following, we define the comparison and aggregation operators on $\hat{S}$.

3.2.1 *Comparison operators*

The comparison of terms is carried out according to the ordinary lexicographic order of $\hat{S}$, i.e., $\forall \hat{s_k}, \hat{s_l} \in \hat{S}$ if $k < l \Leftrightarrow \hat{s_k} < \hat{s_l}$. Following, we define the comparison operators.

Definition 7 $\forall \hat{s_k}, \hat{s_l} \in \hat{S}$ we define the maximization operator as following:

$$\max(\hat{s}_k, \hat{s}_l) = \hat{s}_l \Leftrightarrow \hat{s}_k < \hat{s}_l.$$

We define the minimization operator as following:

$$\min(\hat{s}_k, \hat{s}_l) = \hat{s}_k \Leftrightarrow \hat{s}_k < \hat{s}_l.$$

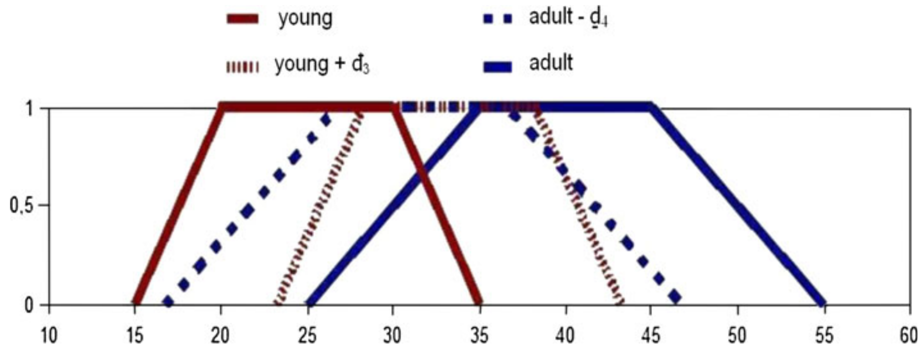


Fig. 5 Maximum and minimum translated terms young and adult

3.2.2 Aggregation operators

The aggregation of information consists of obtaining a value that summarizes a set of values. In order to obtain a more precise result, we obtain this value from $\hat{S}$. It is not possible to define the usual negation operator over $\hat{S}$, due to unbalanced information contained. We proceed to define the aggregation operator based on *ordered weighted averaging* (OWA) and *linguistic ordered weighted averaging* (LOWA) operators (Herrera et al. 1996; Yager 1994).

Let $A = \{a_j\}$, $j \in \{1 \ldots n\}$, $a_j \in S$ be, a set of terms to aggregate; $W = \{w_j\}$, $j \in \{1 \ldots n\}$, $w_j \in [0, 1]$ be their associated weights; and B be the associated ordered term vector. Each element $b_i \in B$ is the i-th largest term in the collection ordered vector $\{a_1, \ldots, a_n\}$. Besides, let the fuzzy operator ε: $S \times S \rightarrow [0, 1]$ be such that $\forall s_i, s_j \in S$, $\varepsilon(s_i, s_j)$ represents the *fuzzy degree of equality of* s_i *over* s_j. There are several possible ways to define the operator ε over A and B (two trapezoidal possibility distributions): possibility and necessity theory (see Table 2), the subjective criterion of some decision maker (Carrasco et al. 2001), etc.

Table 2 Examples of definition of the operator ε

Possibility operator definition ε	Necessity operator definition ε
= sup $_{d \in U}$ min (A(d), B(d)) where U is the domain of A, B. A(d) is the degree of the possibility for d ∈ U in the distribution A (see Fig. 2)	= inf $_{d \in U}$ max (1 − A(d), B(d)) where U is the domain of A, B. A(d) is the degree of the possibility for d ∈ U in the distribution A

Before defining the aggregation operators, we define the following:

Definition 8 We define the *fuzzy degree of equality of* $\hat{s} \in \hat{S}$ *over an ordered set of terms B weighted by W based on an operator* ε:

$$\Theta^{\text{'OWA'}}(\hat{s}, B, W) = \sum_{j=1}^{n} \varepsilon(\hat{s}, b_j) \times w_j$$

Now we proceed to define the aggregation operator:

Definition 9 Let be $\Theta^{\text{'SupOWA'}}(\hat{S}, B, W) = \text{Sup}\{\Theta^{\text{'OWA'}}(\hat{s}_i, B, W), \forall i \in \{1 \ldots m\}\}$. We define the *average over an ordered set of terms B weighted by W based on an operator* ε *respect to* $\hat{S}$ as following:

$$\Theta^{\text{,LOWA,}}(\hat{s}, B, W) = \begin{cases} \hat{s}_j, \quad \text{if } \exists \hat{s}_j, \hat{s}_k \in \hat{s}, j, k \in \{1 \ldots n\} \text{ and } j \neq k / \Theta^{\text{,OWA,}}(\hat{s}_j, BW) \\ \quad = \Theta^{\text{,OWA,}}(\hat{s}_k, B, W) = \Theta^{\text{,SupOWA,}}(\hat{s}, B, W) \text{ and } |d_j| < |d_k| \\ \hat{s} \in \hat{S} / \Theta^{\text{,OWA,}}(\hat{s}, B, W) = \Theta^{\text{,SupOWA,}}(\hat{S}, B, W), \quad \text{otherwise} \end{cases}$$

If $w_j = 1/n$, $\forall j \in \{1 \ldots n\}$ we call to operator $\Theta^{\text{'LOWA'}}(\hat{S}, B, W)$ as quasiarithmetic average over a set of terms A based on an operator ε respect to $\hat{S}$ and we symbolise this definition as $\Theta^{\text{'AVG'}}(\hat{S}, A)$.

We call *the degree of representativeness* of $\Theta^{\text{'LOWA'}}(\hat{S}, B, W)$ operator to a value in [0, 1] defined as:

$$\Theta^{\text{,RepOWA,}}\left(\hat{S}, B, W\right) = \Theta^{\text{,SupOWA,}}\left(\hat{S}, B, W\right) / \sum_{j=1}^{n} w_j$$

Therefore, if the degree of representativeness is the same for more than one term, we choose the term with less semantic translation (in absolute value) in order to define the operator $\Theta^{\text{'LOWA'}}$. This degree should be close to the value 1 for an acceptable representativeness of the chosen term $\Theta^{\text{'LOWA'}}(\hat{S}, B, W)$.

Example 2 Let S be the set of terms defined in the Example 1; let $A = \{$*baby*, *child*, *teenager*, *teenager*,

teenager} be the set of terms to aggregate; and $W = \{0.2, 0.2, 0.2, 0.2, 0.2\}$, i.e., all the terms of A have the same associated weight. Here, we show the solution of this simple aggregation problem by means of the three computational models we have just reviewed in Sect. 2.2 and with the new model FMST:

- Solution based on the Extension Principle (Bonissone and Decker 1986): A linguistic aggregation operator based on the principle acts according to: $S^n \xrightarrow{\Im} F(\Re) \xrightarrow{\text{app1}(\cdot)} S$, where S^n symbolises the n Cartesian product of S, $\Im$ is an aggregation operator based on the extension principle, $F(\Re)$ the set of fuzzy sets over the set of real numbers, and $\text{app}_1(\bullet)$ is a linguistic approximation function that returns a label from the linguistic term set S whose meaning is the closest to the obtained unlabeled fuzzy number. We select the arithmetic mean as the operator $\Im$ and we obtain the fuzzy trapezoidal set $F(\Re) = [7, 8.4, 12.2, 14]$. We apply a linguistic approximation process based on the Euclidean distance and we obtain the value $app_1(\bullet) = s_2 = teenager$.
- Solution based on the ordinal linguistic computational model (Herrera et al. 1996; Delgado et al. 1993): It makes aggregations on the indexes of the ordered linguistic labels according to: $S^n \xrightarrow{C} [0, g] \xrightarrow{\text{app2}(\cdot)} \{0, \ldots, g\} \rightarrow S$, where C is a symbolic linguistic aggregation operator, $\text{app}_2(\bullet)$ is an approximation function used to obtain an index $\{0, \ldots, g\}$ associated to a term in S from a value in $[0, g]$. The operator C we shall use is the convex combination (Delgado et al. 1993). The result of this operator applied on the set A is 1.4. Finally, using as $\text{app}_2(\bullet)$ function the usual round operation, we obtain the value 1, i.e., the term chosen is $s_1 = child$.
- Solution based on the 2-tuple fuzzy linguistic approach (Herrera and Martínez 2000): This method eliminates the loss of information of the previous fuzzy linguistic approaches. However, this method is not applicable with trapezoidal membership functions.
- Solution based on the FMST: As operator σ we will use the possibly operator shown in Table 1; the threshold χ will be 0.75; and h the value 10, in order to obtain the representation model (Definition 6), i.e. the set $\hat{S}$. As operator ε we are going to use the *possibly operator* defined in Table 2, in order to obtain the aggregation operator, showed in Definition 9. Therefore, the quasiarithmetic average over the set of terms A is: $\Theta^{\text{'AVG'}}(\hat{S}, A) = teenager{-}2.2$ and the degree of representativeness of this operator is: $\Theta^{\text{'RepOWA'}}(\hat{S}, B, W) = 0.70$. Therefore, we can say that the quasiarithmetic average over the set of terms A is "2.2 years to *teenager*". This aggregation value has an acceptable representativeness because is close to the value 1.

3.3 Using the FMST model to heterogeneous data management

The FMST computing model can be used to other more heterogeneous contexts, if the definition of the operators σ and ε operate on that type of data. If these operators are defined on any trapezoidal membership function, then we can consider that the set of terms to aggregate $A = \{a_j\}, j \in \{1 \ldots n\}$, $a_j = [\alpha_j, \beta_j, \gamma_j, \delta_j]$. Therefore, the model is directly usable to the following domains that can also be expressed with this type of trapezoidal functions:

- trapezoidal membership function $[\alpha, \beta, \gamma, \delta]$ concerning to *linguistic terms* for S, or other semantic sets for the same terms of S, or even, for other different terms,
- *interval* $[\alpha, \delta]$ with a semantic significance "between α and δ", it can be defined as a trapezoidal function $[\alpha, \alpha, \delta, \delta]$,
- *approximate value* $\alpha \pm m$ with a semantic significance "approximately α with margin $\pm m$", it can be defined with a trapezoidal function $[\alpha - m, \alpha, \alpha, \alpha + m]$,
- *crisp value* α, it can be defined as a trapezoidal function $[\alpha, \alpha, \alpha, \alpha]$,
- missing or undefined values defined as trapezoidal function, for example see (Umano and Fukami 1994): *Unknown defined* as [1, 1, 1, 1], or *Undefined* defined as [0, 0, 0, 0] (see Fig. 6).

Example 3 Let S be the set of terms defined in the Example 1; Let $A = \{child, teenager, [10, 12], 8 \pm 2, 14, unknown\}$ be the set of terms to aggregate; and $W = \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$, i.e., all the terms of A have the same associated weight. The three computational models we have reviewed in Sect. 2.2 are only applicable to linguistic labels. Therefore, to solve this aggregation problem we will only use the FMST. Using the same representation and computational models selected in the Example 2 we have that the quasiarithmetic average over the set of terms A is:

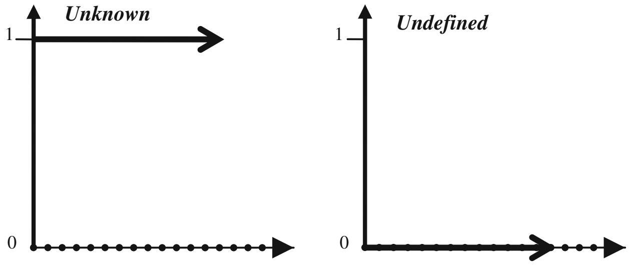


Fig. 6 Unknown and undefined values

$\Theta^{\text{'AVG'}}(\hat{S}, A) = child + 2.6$ and the degree of representativeness of this average is: $\Theta^{\text{'RepOWA'}}(\hat{S}, B, W) = 0.74$. Therefore, we can say that the quasiarithmetic average over the set of terms A is "2.6 years past to *child*". This aggregation value has an acceptable representativeness because is close to the value 1.

3.4 Data summarization using the FMST

Let Y be a set of objects in a database to summarize regarding a set C of attributes that characterize such objects (see Sect. 2.3). Now, we consider that each attribute C_j is defined on a numeric domain as a crisp or even a fuzzy attribute (using the trapezoidal membership functions format explained in the previous section).

Let S_{Cj} be a linguistic term set and let Q_{Cj} be a linguistic quantifier (see Sect. 2.3). In order to obtain the label s_{Cj} that best summarizes the set $C_j(Y)$ and the truth value T_{Cj} using the FMST, the following procedure is proposed:

(1) *Definition of the representation model of the FMST.* We define the set $\hat{S}_{Cj}$ from the set S_{Cj} using the context-free grammar G (Definition 6) choosing an operator σ, a threshold χ and value of discretization h. We consider that this new set $\hat{S}_{Cj}$ has an easy linguistic interpretability (similar to S_{Cj}) and a high precision (depending on the value h).

(2) *Definition of the computational model of the FMST.* We define the operators $\Theta^{\text{'OWA'}}$ and $\Theta^{\text{'LOWA'}}$ (Definition 9) choosing an operator ε.

(3) *Calculation of the label $s_{Cj} \in \hat{S}_{Cj}$, that best summarizes the set $C_j(Y)$ according to the FMST defined.* We choose W_{Cj} as:

$$W_{Cj} = \begin{cases} \{w_i\}/w_i = 1/n, \quad \forall i \in \{1\ldots n\} & \text{if } Q_{Cj} \text{ is a relative quantive} \\ \{w_i\}/w_i = 1, \quad \forall_i \in \{1\ldots n\} & \text{if } Q_{Cj} \text{ is an absolute quantity} \end{cases}$$

And then, we calculate the label that best summarizes the set $C_j(Y)$ as:

$$s_{Cj} = \Theta^{\text{'LOWA'}}\left(\hat{S}_{Cj}, C_j(Y), W_{Cj}\right)$$

(4) *Calculation of the value q_{Cj} an indication (relative or absolute) of the number of pieces of data that satisfy such label s_{Cj}.*

$$q_{Cj} = \Theta^{\text{'OWA'}}\left(s_{Cj}, C_j(Y), W_{Cj}\right)$$

(5) *Calculation of the truth value T_{Cj}* as the membership of q_{Cj} in the proposed quantity in agreement:

$$T_{Cj} = Q_{Cj}\left(q_{Cj}\right)$$

Example 4 Let S be the set of terms defined in the Example 1 and let A and W be defined as in the Example 3. Let $Y = \{y_1, \ldots, y_n\}$ be a set of n records in a database corresponding to a set of hotel guests; Let $C = \{C_1, \ldots, C_r\}$ be the set of attributes of such hotel guests where C_r is the "age of the person". Let the set be $\{C_j(y_i)\} = A$, $\forall i \in \{1\ldots n\}$. Let S_{Cj} be with $S_{Cj} = S$ and let Q_{Cj} be the linguistic quantifier *Most* showed in Fig. 7, see (Galindo et al. 2008) for more information. In order to obtain the label s_{Cj} that best summarizes the set $C_j(Y)$ and the truth value T_{Cj} we use the data summarization model proposed above:

(1) *Definition of the representation model of the FMST.* We define the set $\hat{S}_{Cj}$ from the set S_{Cj} using the context-free grammar G (Definition 6) choosing the *possibly operator* shown in Table 1 for σ, the threshold $\chi = 0.75$ and the value of discretization $h = 10$.

(2) *Definition of the computational model of the FMST.* We define the operators $\Theta^{\text{'OWA'}}$ and $\Theta^{\text{'LOWA'}}$ (Definition 9) choosing the possibly operator defined in Table 2 for ε.

(3) *Calculation of the label $s_{Cj} \in \hat{S}_{Cj}$, that best summarizes the set $C_j(Y)$ according to the FMST defined.* We choose

$$W_{Cj} = \{w_i\}/w_i = 1/n, \quad \forall i \in \{1\ldots n\}$$

because $Q_{Cj} = Most$ is a relative quantity. And then, we calculate the label that best summarizes the set $C_j(Y_k)$ as:

$$s_{Cj} = \Theta^{\text{'LOWA'}}\left(\hat{S}_{Cj}, C_j(Y), W_{Cj}\right) = child + 2.6$$

(4) *Calculation of the value q_{Cj} a relative indication of the number of pieces of data that satisfy such label s_{Cj}:*

$$q_{Cj} = \Theta^{\text{'OWA'}}\left(s_{Cj}, C_j(Y), W_C\right) = 0.74$$

(5) *Calculation of the truth value T_{Cjk}* as the membership of q_{Cjk} in the proposed quantity in agreement:

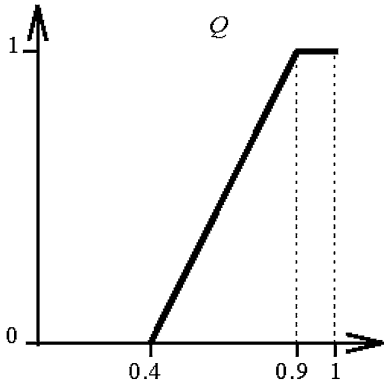


Fig. 7 Linguistic quantifier $Q = Most$

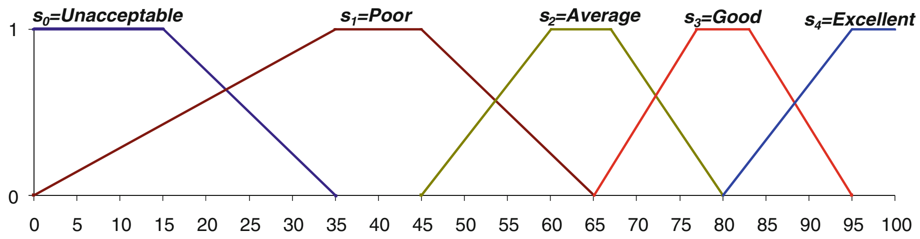
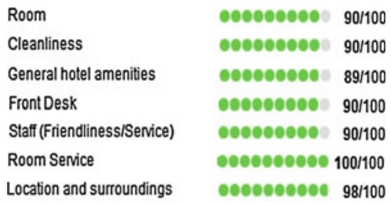


Fig. 8 Membership functions defined for set *S*



Room	90/100
Cleanliness	90/100
General hotel amenities	89/100
Front Desk	90/100
Staff (Friendliness/Service)	90/100
Room Service	100/100
Location and surroundings	98/100

Fig. 9 Example of extraction of ratings from crisp values included in website (Trivago 2011)

$$T_{Cj} = Q_{Cj}\left(q_{Cj}\right) = Q(0.74) = 0.68$$

Therefore, the linguistic summary may be exemplified by "T_{Cj} (most of hotel guests are 2.6 years past to *child*) = 0.68".

4 Applying the linguistic summary of data based on the FMST to tourism

In this section we show the application of the linguistic summary of data, defined in the above section, as a tool which will be part of the architecture of the opinion aggregation showed in the Fig. 1, with the aim of aggregating (over the month dimension) heterogeneous data from various tourism websites in a database. The final goal is to determine the overall opinion of the web community on some high end hotels expressed by means of numerical values, natural language, etc. The selected web pages for data extraction are (Atrapalo 2011; Booking 2011; eDreams 2011; Expedia 2011; TripAdvisor 2011; Trivago 2011). We show an example of the application using the high end hotels located in Granada (Spain) from several websites during the year 2009. Besides, we show examples of analyses that a data analyst can make using the easy linguistic interpretability and the high precision of the model.

We have the following framework in the problem we are trying to solve:

- $Y = \{y_1, \ldots, y_n\}$ is a set of *n* opinions included in the selected web pages for data extraction and obtained after the collect step (showed in the Fig. 1) for each hotel belongs to the set $H = \{hotel_k\}$, with $k \in \{1 \ldots s\}$, in a specific interval of dates corresponding to a higher hierarchy of the time (week, month, etc.) symbolised as *date*.
- $C = \{C_1, \ldots, C_r\}$, with $r = 6$, is a set of fact attributes characterizing such opinions. Next, we explain more precisely the fact attributes: C_1 = *Staff:* in this category we include the room service, the receptionist, etc.; C_2 = *Cleanliness*: especially about the room, including the bathroom; C_3 = *Comfort*: usually going to be heavily influenced by the quality of the bed and the noise level in the room; C_4 = *Location*: reflects on how well the hotel's location is; C_5 = *Price_quality*: includes the price quality ratio of the facilities; C_6 = *Add_Characteristics*: general hotel features not covered in other attributes: design, decor, services, facilities, etc.

Our objective is the aggregation of the data, which is represented in the table: $t_{BD}(date\_id, hotel\_id, C_1, \ldots, C_r, T_{C1}, \ldots, T_{Cr})$ where *id_date* is a temporal attribute with value *date* for the current extraction; *hotel_id* is the identification of the hotel in the set *H*, i.e. $hotel_k$; and C_j and T_{Cj} with $j \in \{1 \ldots r\}$, are the aggregated features of the data sources using the data summarization model and their truth value, respectively, (see Sect. 3.4) for the time *date* and the hotel $hotel_k$.

Therefore, each of these attributes C_j is going to be expressed linguistically by means of the FMST after the aggregation process. With this purpose, the set of primary terms $S = \{s_i\}$, $i \in \{0 \ldots g\}$, with $g = 4$, is defined for each one the attributes, i.e. $S = S_{C1} = \ldots = S_{Cr}$, with the following values: s_0 = *Unacceptable*, s_1 = *Poor*, s_2 = *Average*, s_3 = *Good* and s_4 = *Excellent*. Figure 8 shows the semantic of each one, based on the criterion of expert users and the domain is [0, 100] (as we show below, this domain is the most accurate in the selected web pages). As can be seen, the set of terms are unbalanced. With that, experts have tried to express a higher resolution in the top of the scores because we are dealing with high end hotels, where users are supposed to be very demanding. Besides, we use the relative linguistic quantifier Q = *Most* showed in the Example 4 for each one of attributes, i.e. $Q = Q_{C1} = \ldots = Q_{Cr}$.

Due to the heterogeneous information of the websites, for each one of the facts attributes to obtain, it has been decided to extract information of the following typology:

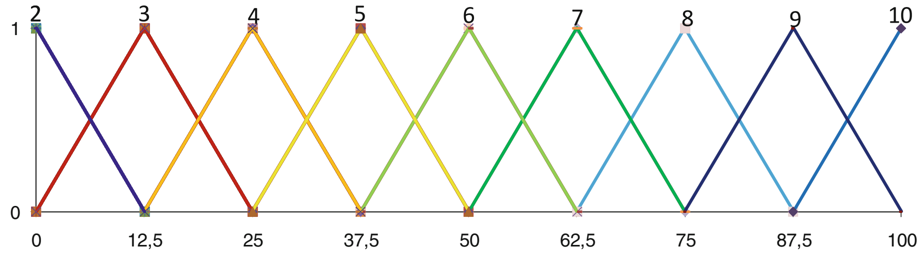


Fig. 10 Membership functions defined for approximate values in the range from 2 to 10

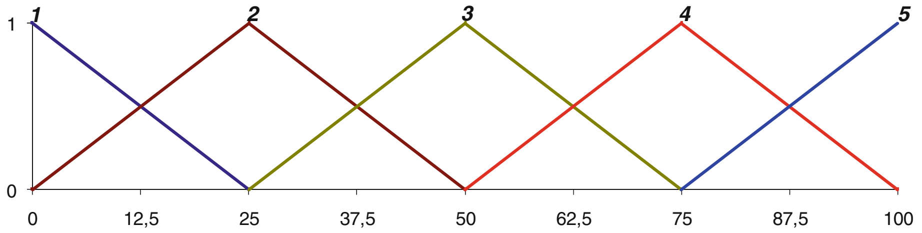


Fig. 11 Membership functions defined for approximate values in the range from 1 to 5

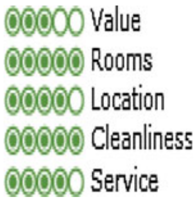
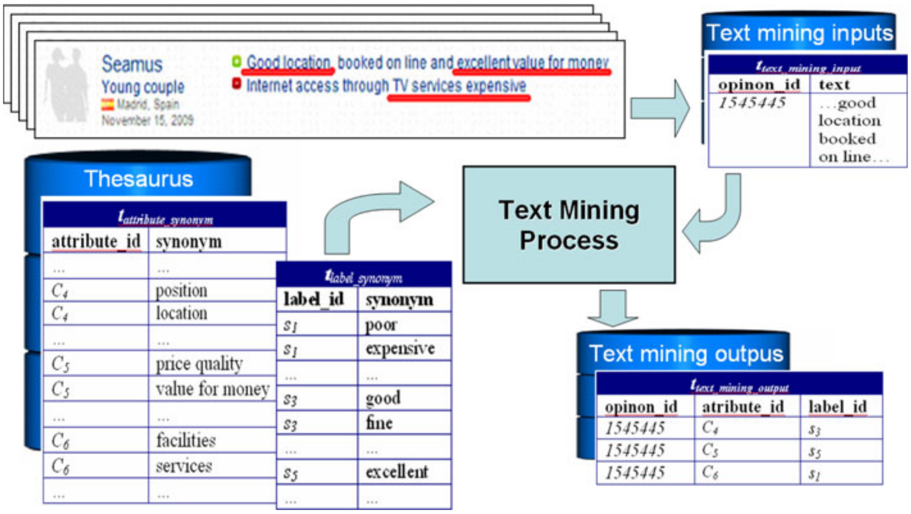

Value
Rooms
Location
Cleanliness
Service

Fig. 12 Example of extraction of ratings from approximate values included in website (TripAdvisor 2011)

- *Crisp values*. In the only site, of those examined, with precise numerical ratings of customers (Trivago 2011), the scores are integer in [0, 100] as shown in Fig. 9. Therefore, there will be no changes of scale for completion of destination attributes: *staff* will correspond to the average between "Front Desk", "Staff" and "Room Service"; *comfort* corresponds to "Room" value; rest is obvious.
- *Approximate values*. Certain pages ask the user to rate the hotels, in the range from 1 to 5 or 2 to 10 without decimals. Since the score that can give the users is far less accurate than the previously described, we will assume this information as approximate values according to Figs. 10 and 11. In the website (TripAdvisor 2011) we have an example of this type of rating (Fig. 12). It is evident the relationship between the attributes in the website and the attributes in our database except for *Add_Characteristics* is not represented on this site as a score explicitly. Such feature will be obtained by searching in the text opinions of the users.
- *Linguistic terms* (*labels*). If the page does not have any numerical value, we decide to get the value of the feature of the hotel from the textual opinions of the clients and we expressed it as a linguistic label. Given the heterogeneity of users that exists, we decide to use a unique semantics for these labels, represented in the Fig. 8. We obtain the pair of features (attribute) and ratings (label) using the text mining schema showed in Fig. 13. Therefore, these processes corresponding to the *Identify* and *Classify steps* showed in the Fig. 1. Oracle Text© (Dixon 2001; Shea 2008) has been used as a tool for the text mining process to manage comments in English and Spanish. In particular, we use the *contains* operator (Shea 2008) in combination with the *near* operator (Shea 2008) to return a score based on the proximity of the two terms searched (characteristic and rating). In order to search both terms, we use some specific thesaurus of terms that contain: synonyms (including formal and informal terms and abbreviators); higher-level terms; and words that have the same root as the specified term [using the *stem* operator (Shea 2008)]. Besides, we use helpful operator for finding more accurate results when there are frequent misspellings in the opinion text: words that sound like the specified terms [*soundex* operator (Shea 2008)] and words that are spelled similarly to the specified terms [*fuzzy* operator (Shea 2008)]. Besides, in the Fig. 13 we show an example of extraction of ratings included in the website (Booking 2011): *Location as Good*, *Price_quality as Excellent*, and the *Add_Characteristics* as *Poor*.
- *Unknown value*. If no one of the previous values has been identified.

The information extracted from web pages is stored in a temporary table t_{TEMP}(*opinion_id, hotel_id*, C_{11}, …, C_{14}, …, C_{r1}, …, C_{r4}) where *opinion_id* is a unique identifier for each review of the hotel (*hotel_id*) in the specific time extracted. Each one of the facts C_j with $j \in \{1 \ldots r\}$, will be represented as a trapezoidal possibility function [C_{j1}, C_{j2}, C_{j3}, C_{j4}] expressing the values that were just comment and



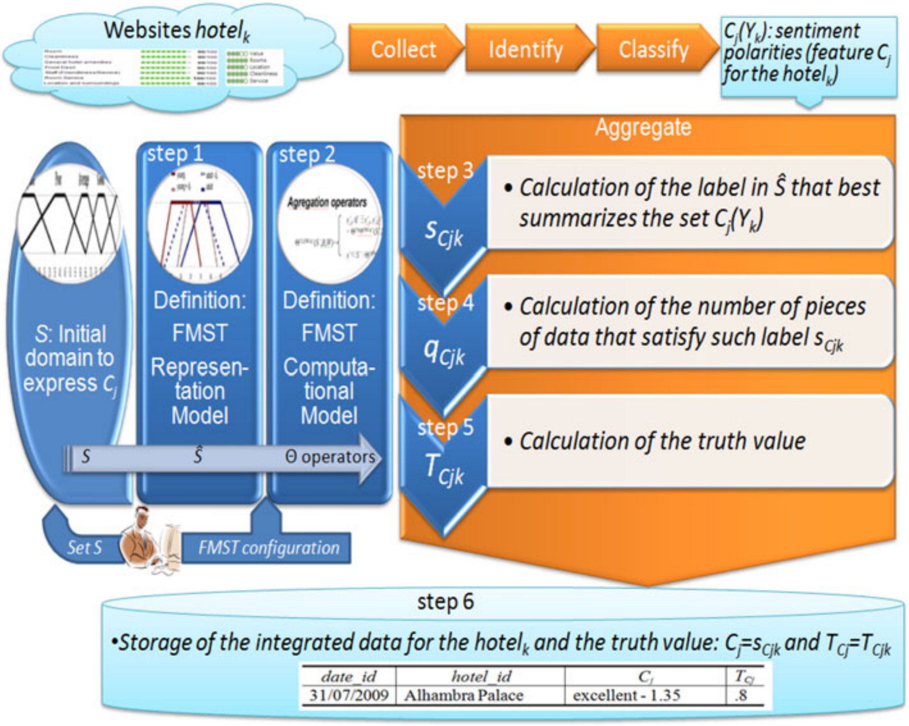
$t_{text\_mining\_input}$	
opinon_id	text
1545445	…good location booked on line…

$t_{attribute\_synonym}$	
attribute_id	synonym
…	…
C_4	position
C_4	location
…	…
C_5	price quality
C_5	value for money
…	…
C_6	facilities
C_6	services
…	…

$t_{label\_synonym}$	
label_id	synonym
s_1	poor
s_1	expensive
…	…
s_3	good
s_3	fine
…	…
s_5	excellent
…	…

$t_{text\_mining\_output}$		
opinon_id	atribute_id	label_id
1545445	C_4	s_3
1545445	C_5	s_5
1545445	C_6	s_1

Fig. 13 Text mining architecture for extraction of ratings from textual opinions. Example of extraction from ratings included in the website (Booking 2011)



- Calculation of the label in $\hat{S}$ that best summarizes the set $C_j(Y_k)$
- Calculation of the number of pieces of data that satisfy such label s_{Cjk}
- Calculation of the truth value

•Storage of the integrated data for the $hotel_k$ and the truth value: $C_j=s_{Cjk}$ and $T_{Cj}=T_{Cjk}$

date_id	hotel_id	C_j	T_{Cj}
31/07/2009	Alhambra Palace	excellent - 1.35	.8

Fig. 14 Procedure to obtain the attributes C_j and T_{Cj} of the table t_{BD} for a $hotel_k$

according to specifications that have been defined in Sect. 3.3. It is going to be named as $C_j(Y_k) = \{[C_{j1}, C_{j2}, C_{j3}, C_{j4}]/hotel\_id = hotel_k\}$ the set of the values that the column C_j store for the hotel $hotel_k$ in the table t_{TEMP}.

In order to obtain the attributes C_j and T_{Cj} of the table t_{BD} using the summarization model here proposed (Sect. 3.4), for each $hotel_k$ with $k \in \{1 \ldots s\}$ we follow the next procedure (see Fig. 14) integrated into the opinion aggregation architecture (shown in Fig. 1):

(1) *Definition of the representation model of the FMST.* In order to increase the accuracy of the final domain of the facts attributes, the set $\hat{S}$ is obtained automatically by means of the grammar showed in Definition 6. As operator σ we will use the possibly operator shown in Table 1. The threshold χ will be 0.75 and h the value 100.

(2) *Definition of the computational model of the FMST.* As operator ε we are going to use the possibly operator defined in Table 2, in order to define the $\Theta^{\text{'OWA'}}$ and $\Theta^{\text{'LOWA'}}$ operators (Definition 9).

(3) *Calculation of the label $s_{Cj} \in \hat{S}_{Cj}$, that best summarizes the set $C_j(Y_k)$ according to the FMST defined.* We choose

$$W_{Cjk} = \{w_i\}/w_i = 1/n, \quad \forall i \in \{1 \ldots n\}$$

because $Q_{Cj} = Most$ is a relative quantity. And then, we calculate the label that best summarizes the set $C_j(Y_k)$ as:

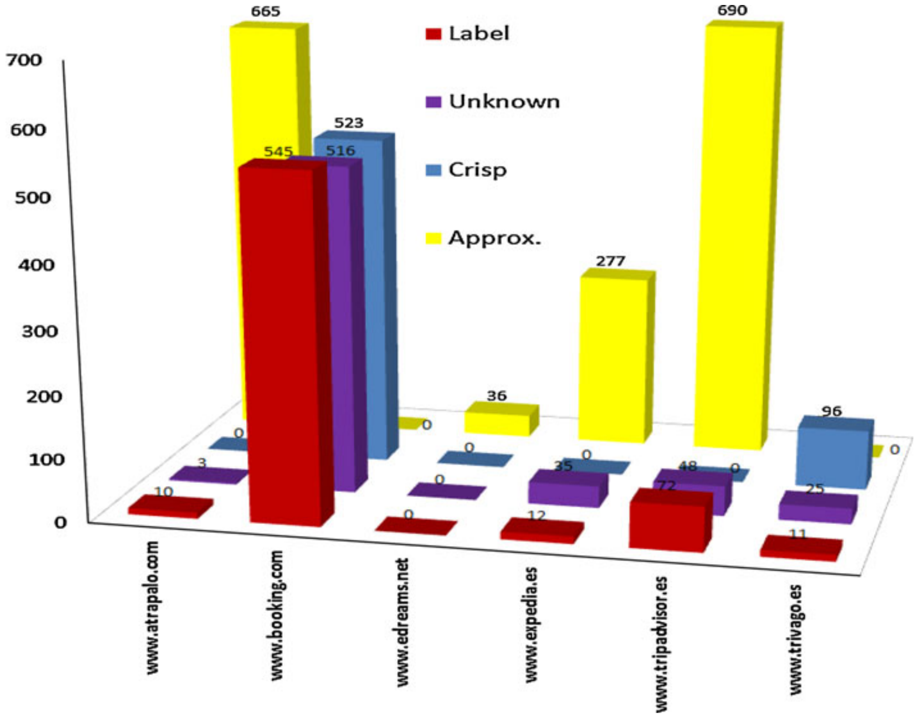


Fig. 15 Statistical outcome of the aggregation process for high end hotels of Granada during the year 2009

Table 3 Table t_{BD} for November of 2009 for all hotels

Hotel_id	C_1	C_2	C_3	C_5	C_4	C_6	T_{C1}	T_{C2}	T_{C3}	T_{C4}	T_{C5}	T_{C6}
Abades Nevada Palace	Good +2.00	Excellent −4.81	Excellent −3.50	Excellent	Poor +15.38	Excellent −7.88	0.6	0.7	0.7	0.6	0.7	0.6
Alhambra Palace	Excellent −4.81	Excellent −3.50	Good +6.50	Good −2.50	Excellent	Good +4.25	0.8	0.9	0.8	0.7	0.9	0.8
Granada Center	Good +6.50	Good −2.50	Good −1.38	Good +2.00	Good −2.50	Good −2.50	0.8	0.8	0.7	0.8	0.9	0.7
Ma Nazaries	Good +4.25	Excellent −0.44	Good +7.63	Excellent −0.44	Good −4.75	Good +5.38	0.6	0.6	0.7	0.7	0.8	0.7
Palacio de los Patos	Excellent	Excellent	Excellent −0.44	Excellent	Excellent	Excellent −4.81	0.6	0.8	0.7	0.7	0.7	0.6
Santos Saray	Excellent	Excellent −0.44	Excellent −4.81	Good +7.63	Good −2.50	Good +4.25	0.9	0.9	0.8	0.6	0.8	0.7

Table 4 Extract of the table t_{BD} for a certain hotel

date_id	*hotel_id*	C_1	C_2	T_{C1}	T_{C2}
30/11/2009	Alhambra Palace	Excellent −4.81	Excellent −3.50	0.8	0.8
31/10/2009	Alhambra Palace	Excellent −3.75	Excellent −3.33	0.7	0.8
30/09/2009	Alhambra Palace	Excellent −3.40	Excellent −3.70	0.9	0.8
31/08/2009	Alhambra Palace	Excellent −0.32	Excellent −0.66	0.8	0.9
31/07/2009	Alhambra Palace	Excellent −1.35	Excellent −1.75	0.8	0.7

$$s_{Cjk} = \Theta^{,\mathrm{LOWA},}\left(\hat{S}_{Cj}, C_j(Y_k), W_{Cjk}\right)$$

(4) *Calculation of the value q_{Cjk} a relative indication of the number of pieces of data that satisfy such label s_{Cjk}.*

$$q_{Cjk} = \Theta^{,\mathrm{OWA},}\left(s_{Cjk}, C_j(Y_k), W_{Ck}\right)$$

(5) *Calculation of the truth value T_{Cjk}* as the membership of q_{Cjk} in the proposed quantity in agreement:

$$T_{Cjk} = Q_{Cj}\left(q_{Cjk}\right)$$

(6) *Storage of the semantic integrated data of the hotel hotel$_k$ during the time "date"*. We insert into the table t_{BD} a tuple with the values for the columns: $date\_id = date$, $hotel\_id = hotel_k$, $C_1 = s_{C1k}$, …, $C_r = s_{Crk}$, $T_{C1} = T_{C1k}$, …, $T_{Cr} = T_{Crk}$.

In the following, we show an example of the application, using as H a set of high end hotels located in Granada (Spain): $H =$ {"Palacio de los Patos", "MA Nazaries", "Alhambra Palace", "Abades Nevada Palace", "Santos Saray", "Granada Center"}. The aggregation process has been made from the above mentioned websites during the year 2009, for monthly periods, according to the architecture explained in this section to obtain the table t_{BD}. Figure 15 shows statistical information result of this process.

With this information, inserted in the table t_{BD}, the user can make several analyses using the easy linguistic interpretability and the high precision of the model according to the most opinions and with an acceptable level of the truth value, such as:

- *Getting the best hotel for a period of time according to their characteristics analyzed*. In Table 3 is possible to verify for the diverse characteristics the major or minor excellence of the hotels during a certain period (November of 2009). Thus, we can conclude that the "Palacio de los Patos" hotel is the best hotel belong to H with a rating of *excellent* for all characteristics analyzed except the additional characteristics (*Add_Characteristics* attribute) with a rating of "4.81 points (over 100) to *excellent*" and *comfort* with a rating of "0.41 points to *excellent*", i.e., practically *excellent*. This assessment agrees with objective data such as the several awards received by the hotel: it has been included on the Condé Nast Traveller (2011) Gold List of "The Best Hotels in the World" in the year 2006; the German edition of the prestigious GEO magazine (GEO Saison 2011) has acknowledged this property as the second "Best Design Hotel in Europe" in 2006; the international luxury travel guide Conde Nast Johansens (2011) has awarded it as the "Most Excellent European Hotel for Design & Innovation 2008"; the hotel also received two other important prizes at the SLEEP 06 European Hotel Design Awards (TheSleepEvent 2011): the "2006 Hotel Design of the Year" and the "Best Hotel Architecture" award in the "Conversion" category.
- *Identifying the weaknesses and strengths of the characteristics of every hotel and comparison with regard to the others*. For example, in the Table 3 is easy to study the location of the hotels: "Abades Nevada Palace" as "15.38 past to *poor*", it is a 30-min walk to the centre of city; "MA Nazaries" as "4.75 to *good*", it is a 15-min walk to the centre of city; "Santos Saray" and "Granada Center" as "2.50 to *good*", theses hotels are located in two modern districts adjacent to the downtown; "Palacio de los Patos" and "Alhambra Palace" as *excellent*, the first is placed in the historical centre of the city and the second is very near the most visited monument in Spain (with about 3 million visitors in 2009): the Alhambra.
- *Historical evolutions of the characteristics of the hotels*. For example, Table 4 shows the evolution from July of 2009 to November of 2009 for a certain hotel and the characteristics *staff* and *cleanliness*. The study of these two characteristics, which depend directly on the hotel workers, may be particularly interesting to the hotel management. It is easy to verify the improvement during summer months, probably due to the lower occupation of the hotel, since Granada is not a coastal city with high inland tourism.

5 Conclusions

In this paper we have presented a FMST applied to the linguistic summarization of databases. We have used the new approach as a tool which is included in an opinion aggregation architecture, with the aim of aggregating heterogeneous data from various tourism websites. Therefore, opinions about high end hotels expressed by means of natural language, approximate, numerical and missing values have been integrated in a database. With this information, a data analyst can make several analyses with the benefit of easy linguistic interpretability and a high precision. The characteristics of the model here proposed are not together on any of the existing linguistic models. We are currently focusing on the theoretical study of the properties of the FMST depending on the fuzzy operators chosen (superiority and equality). Besides, we are working on extending the solution proposed here to similar aggregation problems: opinion poll and reputational risk management.

References

Atrapalo (2011) Travel agency and promotion of recreational activities on the internet. http://www.atrapalo.com

Bonissone PP (1982) A fuzzy sets based linguistic approach: theory and applications. In: Gupta MM, Sanchez E (eds) Approximate reasoning in decision analysis. North-Holland, Amsterdam, pp 329–339

Bonissone PP, Decker KS (1986) Selecting uncertainty calculi and granularity: an experiment in trading-off precision and complexity. In: Kanal LH, Lemmer JF (eds) Uncertainty in artificial intelligence. North-Holland, Amsterdam, pp 217–247

Booking (2011) Europe's leading online hotel reservations agency by room nights sold. http://www.booking.com

Bordogna G, Passi G (1993) A fuzzy linguistic approach generalizing boolean information retrieval: a model and its evaluation. J Am Soc Inf Sci 44:70–82

Bordogna G, Passi G (2001) An ordinal information retrieval model. Int J Uncertain Fuzziness Knowl Based Syst 9:63–76

Bouchon-Meunier B, Yao J (1992) Linguistic modifiers and imprecise categories. Int J Intell Syst 7:25–36

Carenini G, Ng RT, Zwart E (2005) Extracting knowledge from evaluative text. In: Proceedings of the 3rd international conference on knowledge. ACM Press, New York, pp 11–18

Carrasco RA, Galindo J, Vila MA (2001) Using artificial neural network to define fuzzy comparators in FSQL with the criterion of some decision-maker. Lect Notes Comput Sci 2085:587–594

Condé Nast Traveller (2011) The luxury travel website of Condé Nast traveller magazine. http://www.cntraveller.com

Delgado M, Verdegay JL, Vila MA (1992) Linguistic decision making models. Int J Intell Syst 7:479–492

Delgado M, Verdegay JL, Vila MA (1993) On aggregation operations of linguistic labels. Int J Intell Syst 8:351–370

Dixon P (2001) Basics of oracle text retrieval. IEEE Data Eng Bull 24(4):11–14

eDreams (2011) Offers the widest selection and the best prices on the market for flights, hotels and vacation packages. http://www.edreams.net

Expedia (2011) Broadest selections of travel products. http://www.expedia.com

Galindo J, Carrasco RA, Almagro AM (2008), Fuzzy quantifiers with and without arguments for databases: definition, implementation and application to fuzzy dependencies. In: Proceedings 12th international conference information processing and management of uncertainty for knowledge-based systems, Malaga, Spain, pp 227–234

George R, Srikanth R (1996) Data summarization using genetic algorithms and fuzzy logic. In: Herrera F, Verdegay JL (eds) Genetic algorithms and soft computing. Physical, Heidelberg, pp 599–611

GEO Saison (2011) A multithematical magazine dedicated to tourism. http://www.geo.de

Herrera F, Martínez L (2000) A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 8(6):746–752

Herrera F, Herrera-Viedma E, Verdegay JL (1995) A sequential selection process in group decision making with linguistic assessment. Inf Sci 85:223–239

Herrera F, Herrera-Viedma E, Verdegay JL (1996) Direct approach processes in group decision making using linguistic OWA operators. Fuzzy Sets Syst 79:175–190

Herrera F, Martínez L, Sánchez PJ (2005) Managing non-homogeneous information in group decision making. Eur J Oper Res 166(1):115–132

Herrera-Viedma E (2001) An information retrieval system with ordinal linguistic weighted queries based on two weighting elements. Int J Uncertain Fuzziness Knowl Based Syst 9:77–88

Herrera-Viedma E, López-Herrera AG, Luque M, Porcel C (2007) A fuzzy linguistic IRS model based on a 2-tuple fuzzy linguistic approach. Int J Uncertain Fuzziness Knowl Based Syst 15:225–250

Hu M, Liu B (2004) Mining opinion features in customer reviews. In: Proceedings of nineteenth national conference on artificial intelligence. San José, California, pp 755–760

Johansens CN (2011) Luxury hotels, spas & venues from Condé Nast Johansens. http://www.johansens.com

Kacprzyk J (1999) An interactive fuzzy logic approach to linguistic data summaries. In: Proceedings 18th international conference of the North American fuzzy information processing society, New York, pp 595–599

Kacprzyk J, Yager RR (2001) Linguistic summaries of data using fuzzy logic. Int J General Syst 30:33–154

Kacprzyk J, Zadrozny S (2000) Computing with words: towards a new generation of linguistic querying and summarization in databases. In: Sincak P, Vašcak J (eds) Quo Vadis computational intelligence? Physica, Heidelberg, pp 144–175

Kacprzyk J, Yager RR, Zadrozny S (2000) A fuzzy logic based approach to linguistic summaries in databases. Int J Appl Math Comput Sci 10:813–834

Klement EP, Mesiar R, Pap E (2000) Triangular Norms. In: Klement EP, Mesiar R (eds) Trends in logic vol 8, Studia Logica Library, Kluwer Academic Publishers, Dordrecht

Ku LW, Liang YT, Chen HH (2006) Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI-2006 Spring symposium on computational approaches to analyzing weblogs. Menlo Park, California, pp 100–107

Laurent A (2003) A new approach for the generation of fuzzy summaries based on fuzzy multidimensional databases. Intell Data Anal 7(2):155–177

Lazzari LL, Mouliá PI, Eriz M (2009) An alternative operationalization of fuzzy consideration set. Application to tourism. In: Proceedings of IFSA/EUSFLAT Conference. Lisboa, Portugal, pp 173–177

Long C, Zhang J, Huang M, Zhu X, Li M, Ma B (2009) Specialized review selection for feature rating estimation. In: Proceedings of the IEEE/WIC/ACM international conference on web intelligence. Milan, Italy, pp 214–221

Miao Q, Li Q, Dai R (2009) Amazing: a sentiment mining and retrieval system. Expert Syst Appl 36(3):7192–7198

Moreno JM, Morales del Castillo JM, Porcel C, Herrera-Viedma E (2010) A quality evaluation methodology for health-related websites based on a 2-tuple fuzzy linguistic approach. Soft Comp 14(8):887–897

Morinaga S, Yamanishi K, Tateishi K, Fukushima T (2002) Mining product reputations on the web. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. Press, New York, pp 341–349

Petrík M (2010) Convex combinations of strict t-norms. Soft Comp 14(10):1053–1057

Ribeiro A, Fresno V, García-Alegre M, Guinea D (2002) A fuzzy system for the web page representation. In: Szczepaniak PS, Segovia J, Kacprzyk J, Zadeh LA (eds) Intelligent exploration of the web. Physica, Heidelberg pp 19–38

Shea C (2008) Oracle text reference, 11g release 1 (11.1) Part Number B28304-03

Tang H, Tan S, Cheng X (2009) A survey on sentiment detection of reviews. Expert Syst Appl 36(7):760–773

TheSleepEvent (2011) The sleep event conference. http://www.thesleepevent.com

TripAdvisor (2011) Branded sites alone make up the most popular and largest travel community in the world. http://www.tripadvisor.es

Trivago (2011) A premiere international online service for travelers seeking advice regarding their travel destinations. http://www.trivago.com

Tsytsarau M, Palpanas T (2010) Mining subjective data on the web. In: Technical report DISI-10-045, Ingegneria e Scienza dell'Informazione. University of Trento, Italy

Umano M, Fukami S (1994) Fuzzy relational algebra for possibility-distribution-fuzzy-relational model of fuzzy data. J Intell Inf Syst 3:7–28

Yager RR (1982) A new approach to the summarization of data. Inf Sci 28:69–86

Yager RR (1991) On linguistic summaries of data. In: Frawley W, Pietsky-Shapiro G (eds) Knowledge discovery in databases. AAAI/MIT Press, Cambridge, pp 347–363

Yager RR (1995) An approach to ordinal decision making. Int J Approx Reas 12(3–4):237–261

Yager RR (1994) On weighted median aggregation. Int J Uncertain Fuzziness Knowl Based Syst 2:101–113

Yager RR (1999) Decision making under uncertainty with ordinal information. Int J Uncertain Fuzziness Knowl Based Syst 7:483–500

Zadeh LA (1975) The concept of a linguistic variable and its applications to approximate reasoning, Pt I, Inf Sci 8:199–249. Pt II, Inf Sci 8:301–357. Pt III, Inf Sci 9:43–80

Zhang J, Kawai Y, Kumamoto T, Tanaka K (2009) A novel visualization method for distinction of web news sentiment. In: Vossen G, Long DDE, Yu JX (eds) LCNS, vol 5802. Springer, Berlin, pp 181–194