# Profile reliability to improve recommendation in social-learning context

Corinne Amel Zayani, Leïla Ghorbel, Ikram Amous, Manel Mezghani, André Péninou, Florence Sèdes

# Profile reliability to improve recommendation in social-learning context

Corinne Amel Zayani, Leila Ghorbel and Ikram Amous
*MIRACL-ISIMS, University of Sfax, Sfax, Tunisia, and*
Manel Mezghanni, André Péninou and Florence Sèdes
*IRIT UMR CNRS 5505, University of Toulouse 3, Toulouse, France*

## Abstract

**Purpose** – Generally, the user requires customized information reflecting his/her current needs and interests that are stored in his/her profile. There are many sources which may provide beneficial information to enrich the user's interests such as his/her social network for recommendation purposes. The proposed approach rests basically on predicting the reliability of the users' profiles which may contain conflictual interests. The paper aims to discuss this issue.

**Design/methodology/approach** – This approach handles conflicts by detecting the reliability of neighbors' profiles of a user. The authors consider that these profiles are dependent on one another as they may contain interests that are enriched from non-reliable profiles. The dependency relationship is determined between profiles, each of which contains interests that are structured based on *k*-means algorithm. This structure takes into consideration not only the evolutionary aspect of interests but also their semantic relationships.

**Findings** – The proposed approach was validated in a social-learning context as evaluations were conducted on learners who are members of Moodle e-learning system and Delicious social network. The quality of the created interest structure is assessed. Then, the result of the profile reliability is evaluated. The obtained results are satisfactory. These results could promote recommendation systems as the selection of interests that are considered of enrichment depends on the reliability of the profiles where they are stored.

**Research limitations/implications** – Some specific limitations are recorded. As the quality of the created interest structure would evolve in order to improve the profile reliability result. In addition, as Delicious is used as a main data source for the learner's interest enrichment, it was necessary to obtain interests from other sources, such as e-recruitement systems.

**Originality/value** – This research is among the pioneer papers to combine the semantic as well as the hierarchical structure of interests and conflict resolution based on a profile reliability approach.

**Keywords** Enrichment, Interests, Reliability, Temperature, *k*-means, Semantic similarity

## 1. Introduction

Recommender systems suggest to users resources (called items) relative to their interests. These systems are based on three basic approaches including collaborative filtering, content-based filtering and knowledge-based recommendation (Felfernig *et al.*, 2014). These approaches suffer from certain limits such as data sparsity and cold start problems (Jadhav and Wankhade, 2016).

Recently, social-based recommendation approaches have emerged to overcome these limits (Kumar *et al.*, 2016; Mezghani *et al.*, 2017; Kalaï *et al.*, 2017; Yu *et al.*, 2018). Generally, these approaches enrich the user's profile, in different contexts (e-commerce, e-learning, etc.), with interests extracted from many sources such as his/her social behavior (Mezghani *et al.*, 2017) (e.g. tagging behavior) or his/her social profiles (Martinez *et al.*, 2014; Kalaï *et al.*, 2017). Particularly, in e-learning context, the learner's profile may contain incomplete or partial data. As a consequence, there is a strong need to enrich the learner's profile in a social-learning context for recommendation purposes (e.g. recommend: relevant pedagogic resources for learning based on the learner's social interests, traineeships, jobs, etc.).

DOI 10.1108/OIR-02-2017-0068

With the evolutionary aspect of social networks, the enrichment of the user's/learner's profile with interests from various social profiles becomes a crucial problem. This problem resides in the fact that interests that are selected for enrichment may be conflictual (out-of-date, duplicate, ambiguous, etc.). Most of the conflict resolution approaches (Li *et al.*, 2016) in different fields such as users' profiling field (Varma *et al.*, 2017; Li *et al.*, 2017) are based on source reliability methods. These methods tend to detect the reliability of sources (profiles) and trustworthy data values (values of interests) in the corresponding sources (Li *et al.*, 2016). We notice that source reliability methods rest on non-organized data namely ignoring the semantic relationship between data and their evolutionary aspect over time.

In this paper, we propose a new approach to resolve conflicts between interests for recommendation purposes. This approach detects reliable neighbors' profiles of a user based on organized profiles. We propose to organize each profile by generating a hierarchical structure of interests that takes into account their semantic relationships as well as their evolutionary aspect over time. This approach was validated, in social-learning context, based on learners who are members of Moodle e-learning system and Delicious social network. Results show that the generated hierarchies improve the results of conflict resolution by predicting the reliability of profiles. These results could promote recommendation systems as the selection of interests that are considered for enrichment depends on the reliability of profiles.

The rest of this paper is organized as follows. In Section 2, some existing studies about users' profile reliability and conflict resolution are presented and discussed. Then, in Section 3, we give an overview of our approach. In Sections 4 and 5, we identify the proposed mechanisms of our approach. In Section 6, we describe the evaluation results. Finally, Section 7 concludes the paper and offers certain prospects for future works.

## 2. Related works

In this section, an overview about reliability of users' profiles and methods for conflict resolution are presented followed by a synthesis.

### *2.1 Reliability of users' profiles*

In literature, the reliability of users' profiles appeared with user profiling approaches (Varma *et al.*, 2017; Li *et al.*, 2017). These approaches consist in extracting, from different sources/profiles, data about a specific user (interests, preferences, goals, background, etc.) and enriching with these data his/her profile. The major challenges of user profiling approaches (Barforoush *et al.*, 2017) are: the reliability of the profiling sources, the data that constitute the profile and the enrichment techniques.

The reliability of the profiling sources consists in detecting the reliability of profiles in order to obtain the required data about a user. These data should be evacuated from conflicts. In order to resolve conflicts, some approaches are based on source reliability methods (Varma *et al.*, 2017; Li *et al.*, 2017). These methods are detailed in the next subsection. Other approaches such as in educative area consider that a profile is reliable only if it is frequently updated (Martinez *et al.*, 2014) or manually notified (Walsh *et al.*, 2013) by the learner.

It is also interesting to consider data that constitute the profile for user profiling. In fact, a profile may contain data with conflicts (irrelevant: duplicate, ambiguous, out-of-date, etc.). In this context, some approaches tend to organize the user's profile in order to maintain relevant data. Generally, the profile organization is carried out mainly by using the machine learning techniques (co-training (Ghorbel *et al.*, 2016), KNN (Xu *et al.*, 2015), *k*-means (Li *et al.*, 2016), etc.). These techniques permit the structuring of the user's data by deleting the out-dated ones.

As the user's data and particularly his/her interests change over time, other studies use the notion of temperature including the interest freshness and popularity (Mezghani *et al.*, 2014) to keep in the profile some popular interests (interests of a wide number of similar users) relative to a specific period of time.

Enrichment consists in adding relevant data in the corresponding profile in order to improve its content. Several studies, in literature, applied three different techniques allowing the validation of the relevance of data (interests). The first technique rests on using the vector space model (Gemmell *et al.*, 2008) that consists in extracting and adding the keywords relative to the user's queries or tags reflecting the user's interests. However, it is likely that the system enriches the user's profile with redundancy, ambiguity and lack of semantics. In order to resolve this problem, other studies proposed the knowledge-based technique (Simpson, 2008).

The second technique considers the semantic relationship between interests and an external semantic dictionaries, such as Wordnet (http://wordnet.princeton.edu). Generally, these dictionaries provide a restrictive set of words and concepts. Thus, the new used concepts and words that do not exist in the dictionary are not considered for enrichment. For this reason, the context-based technique appears to overcome this limit.

The third technique is based not only on semantic dictionary, but also on other knowledge resources such as the interests relative to social profiles (profiles of the user's closest friends or neighbors) (Mezghani *et al.*, 2014). This technique needs to apply a conflict resolution method in order to enrich the user's profile with the most relevant data.

### *2.2 Methods for conflict resolution*

The conflict resolution is made through two methods which are majority voting as well as source reliability (Li *et al.*, 2016).

The majority voting method rests on merging in the corresponding source (profile in our case) data with the highest number of occurrences existing in the other sources. The major shortcoming of this method is that it assumes that all sources providing data are equally reliable (Li *et al.*, 2016). As a matter of fact, the second method emerged in order to estimate the source reliability degrees and infer true data. The sources providing true data will be assigned higher reliability, and data supported by reliable sources will be regarded as true data (Dong *et al.*, 2013).

Source reliability methods are extremely interesting as they lay the ground for certain beneficial applications in different fields including user profiling field. In literature, we find several methods (Li *et al.*, 2016) that are established in order to distinguish reliable and non-reliable sources by inferring their reliability degrees and to derive true data by conducting weighted aggregation (Dong *et al.*, 2013; Pasternack and Roth, 2010; Li *et al.*, 2014; Zhao *et al.*, 2014; Zhang *et al.*, 2016). These methods adopt some characteristics that are summarized in several aspects, such as input data, source reliability and output.

The input aspect describes the pre-processing of the input data which can be duplicated. In order to solve the problem of data duplication, some studies take into account the evolutionary aspect of data over time by considering their freshness values (Sarma *et al.*, 2011; Rekatsinas *et al.*, 2014; Huang *et al.*, 2017; Li *et al.*, 2017; Varma *et al.*, 2017).

Moreover, the input data can be structured (Dong *et al.*, 2013; Pasternack and Roth, 2010; Dong *et al.*, 2014; Varma *et al.*, 2017), unstructured (Yu *et al.*, 2014), categorical (Huang *et al.*, 2017), continuous (Pasternack and Roth, 2010) or heterogeneous (Li *et al.*, 2014; Zhang *et al.*, 2016; Li *et al.*, 2017). For example, in Dong *et al.* (2014), the input data are related to spaces in the world which are represented in a hierarchical structure.

The source reliability aspect describes the used assumptions. The most popular assumption is related to the source dependency. Some studies assume that sources are independent as they do not copy data from each other (Li *et al.*, 2014; Huang *et al.*, 2017).

Other studies assume that sources are dependent on one another (Dong *et al.*, 2013; Sarma *et al.*, 2011; Huang and Wang, 2016; Li *et al.*, 2017). The authors adjust the weight of each source based on the copying relationship among sources. In fact, a source can copy data from non-reliable source (direct copying) and these data can be copied to another source (co-copying, transitive copying).

As for the output aspect, these approaches use either the labeling technique (Pasternack and Roth, 2010), which assigns a label (true or false) to each source or a scoring technique that assigns a score to each source in the form of a probability (Dong *et al.*, 2013; Zhang *et al.*, 2016; Rekatsinas *et al.*, 2017).

### *2.3 Synthesis*

Departing from this state-of-the-art, we can notice that the conflict resolution has whetted the interest of different researchers in different fields, which gave birth to several approaches. The most popular approaches rely on source reliability methods.

Table I presents a comparative study about these approaches. The comparison rests on aspects of source reliability methods: input aspect, assumptions used to estimate the reliability of sources and output.

We notice that despite the efforts provided in some approaches in order to improve the result of source reliability based on unstructured or structured and heterogeneous (categorical and continuous) input data, these approaches suffer from two main shortcomings.

The first limit resides in the fact that some of them have ignored data evolution over time (temperature) in each source (Pasternack and Roth, 2010; Dong *et al.*, 2013, 2014; Li *et al.*, 2014; Huang and Wang, 2016), which improves the result of source reliability in (Yin and Tan, 2011; Yu *et al.*, 2014; Huang *et al.*, 2017; Varma *et al.*, 2017; Li *et al.*, 2017).

The second limit lies, to the best of our knowledge, in the fact that all these approaches do not take into account the semantic relationship between data in sources. Indeed, ignoring the semantic relationship between data may affect the accuracy of these methods because data existing in sources cannot be regarded as separate.

In fact, the learner's profile as a source contains different values, especially those of interests which are semantically dependent on each other in the same period of time. The user's interests can be characterized by their temperature value (Manzat *et al.*, 2010; Mezghani *et al.*, 2014), including their freshness and popularity values that change over time.

Thus, we need to know how far these interests are up-to-date (freshness) and popular and decide whether some of them are irrelevant in order to exclude them from enrichment. As a consequence, the organization of the user's profile by taking into account the semantic relationship between interests and their temperature values is required.

Moreover, we notice that the majority of approaches that are based on source dependency assumption assign scores to sources (Yin and Tan, 2011; Dong *et al.*, 2013; Huang and Wang, 2016; Li *et al.*, 2017). These approaches are motivated to choose the scoring technique because sources have a certain probability (score) to be reliable. However, with the labeling techniques (Pasternack and Roth, 2010; Varma *et al.*, 2017), this information is lost.

In this paper, we propose a users' profile reliability approach to resolve conflicts between interests for recommendation purposes. This profile was validated in social-learning context. In fact, a learner may benefit from his/her social data including data stored in the profiles of his/her friends or neighbors. The originality of this approach resides in the fact that it detects reliable neighbors' profiles of a user/learner by applying some aspects of the source reliability methods based on organized profiles. A user's profile is organized by generating a semantic and hierarchical structure of the user's interests based on *k*-means machine learning algorithm.

**Table I.**
Comparative study of source reliability approaches

| Approaches | Aspects | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Input | | | | Organization | | Assumption | | Output | |
| | Unstructured | Structured | Continuous | Categorical | Semantic | Temperature | Non-dependency | Dependency | Labels | Scores |
| Pasternack and Roth (2010) | – | ✓ | – | ✓ | – | – | ✓ | – | ✓ | – |
| Yin and Tan (2011) | ✓ | ✓ | ✓ | ✓ | – | ✓ | – | ✓ | – | ✓ |
| Dong *et al.* (2013) | – | – | – | ✓ | – | – | – | ✓ | – | ✓ |
| Li *et al.* (2014) | – | ✓ | ✓ | ✓ | – | – | ✓ | – | – | ✓ |
| Dong *et al.* (2014) | ✓ | | – | ✓ | – | – | ✓ | – | – | ✓ |
| Yu *et al.* (2014) | ✓ | – | ✓ | ✓ | – | – | – | ✓ | – | ✓ |
| Huang and Wang (2016) | ✓ | – | ✓ | ✓ | – | – | – | ✓ | – | ✓ |
| Huang *et al.* (2017) | – | ✓ | – | ✓ | – | ✓ | – | ✓ | – | ✓ |
| Varma *et al.* (2017) | ✓ | – | ✓ | ✓ | – | ✓ | ✓ | – | ✓ | – |
| Li *et al.* (2017) | – | ✓ | ✓ | ✓ | – | ✓ | ✓ | – | – | ✓ |
| Our approach | – | ✓ | ✓ | ✓ | ✓ | ✓ | – | ✓ | – | ✓ |

## 3. Overview of the proposed profile reliability approach

In this section, the principle of the proposed approach is first presented. Second, the used factors are exhibited.

### *3.1 Principle*

The user's profile and particularly the user's interest can be enriched from several profiles for recommendation purposes. These profiles may be non-reliable as they may provide interests with conflicts namely false interests. For this reason, our approach attempts to detect the most reliable profiles in order to extract true interests for enrichment.

We denote $\mathcal{P} = \{p_1, ..., p_k, ..., p_n\}$ the set of profiles of the user and the user's neighbors. Each profile includes a set of interests denoted by $I = \{i_1, ..., i_a, ..., i_m\}$. Each interest ($i_a$) in $p_k$, denoted by $p_k(i_a)$, is represented by three elements: a word, its freshness and its popularity.

Our approach rests on aspects of source reliability methods for which we propose improvements. Figure 1 illustrates the integration of these aspects with the mechanisms of the proposed approach as well as the used factors. The latter takes into account the following aspects:

- Input data: most of the proposed research studies are confined to the evolution of data over time (Yin and Tan, 2011; Yu *et al.*, 2014; Huang *et al.*, 2017; Varma *et al.*, 2017; Li *et al.*, 2017), which we judge insufficient to resolve conflicts. Indeed, we find that taking into account the semantic relationship between data may improve the result of the profile reliability. For this reason, in our work, we take into account, on the one hand, the degree of semantic similarity between interests through the semantic factor and on the other hand, their evolutionary characteristics that encompass the values of freshness and popularity. These values are designated by the temperature factor that distinguishes between recent/non-recent and/or popular/non-popular interests. Semantic and temperature factors are used to represent the interests in a semantic and hierarchical structure. For this reason, we propose an organization mechanism which is based on the unsupervised machine learning technique. This technique can automatically affect interests to groups instead of manually assigning labels to all interests that are very numerous. Thus, we are based on the $k$-means algorithm which is the most popular among the unsupervised algorithms.
- Source reliability assumption: most of the proposed research studies are based on the estimation of the dependence between sources. They are based on structured input data by taking into account their evolutionary characteristics uniquely. In our work, we assume that profiles are dependent on one another by taking into account not only

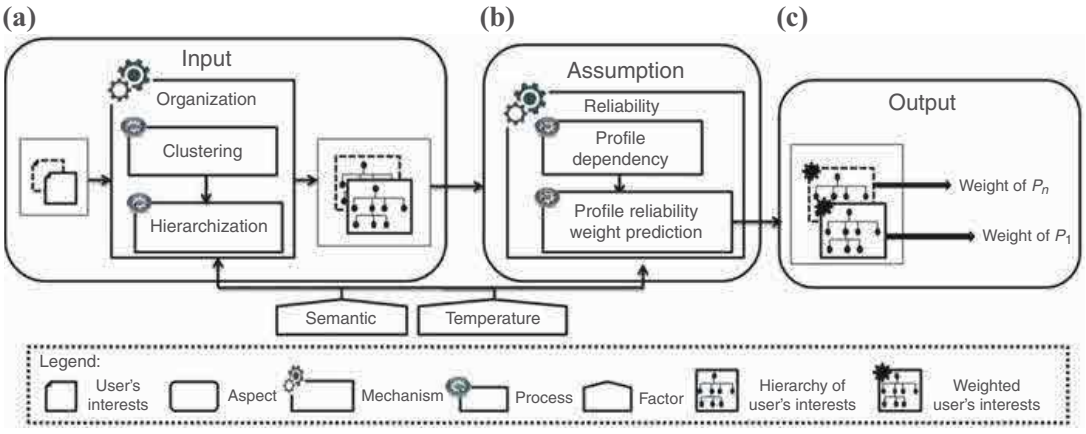



**Figure 1.**
The proposed profile reliability approach

the evolutionary characteristics of interests (temperature factor), but also their semantic relationships (semantic factor). This assumption allows the prediction of the degree of dependency between the profiles based on the semantic and hierarchical structure of the interests of each profile (organization mechanism). Subsequently, it makes possible the prediction of the profile reliability weights (scores). For this reason, we propose a reliability mechanism.

- Output: for this aspect, we propose to assign weights to interests and profiles. For this reason, we assign a weight to each interest with a probability value. This weight is calculated based on the degrees of dependency of profiles and the temperature factor. Subsequently, the interest weights of a profile are aggregated to predict its reliability weight based on the reliability mechanism.

The above mentioned aspects stand for the cornerstone of the reliability mechanism which is in turn based on the organization mechanism by using the semantic and temperature factors.

### *3.2 Semantic factor*

This factor is used in order to measure the similarity between two interests based on semantic similarity measures. There are six measures of semantic similarity (Gomaa and Fahmy, 2013); three of them are based on information content (IC): Resnik (res), Lin (lin) and Jiang and Conrath (jcn). The other three measures are based on Path Length (PL): Leacock and Chodorow (lch), Wu and Palmer (wup) and path.

In our work, the similarity between two interests $i_a$ and $i_b$ is measured through the combination between the IC and PL measures based on the following equation:

$$\text{Similarity}(i_a, i_b) = \alpha \times \text{IC}(i_a, i_b) + \beta \times \text{PL}(i_a, i_b)\,, \tag{1}$$

with $\alpha$ and $\beta$ are the weighting parameters in [0, 1] and $\beta = 1-\alpha$.

Equations (2) and (3) represent, respectively, the average of the similarity values between all IC (res, lin and jcn) measure values and all PL (lch, wup and path) measure values:

$$\text{IC}(i_a, i_b) = \text{Average}(\text{res}(i_a, i_b), \text{lin}(i_a, i_b), \text{jcn}(i_a, i_b))\,, \tag{2}$$

$$\text{PL}(i_a, i_b) = \text{Average}(\text{lch}(i_a, i_b), \text{wup}(i_a, i_b), \text{path}(i_a, i_b))\,. \tag{3}$$

### *3.3 Temperature factor*

Generally, temperature reflects the importance of a resource (document: text, video, image, etc.) for a user (Manzat *et al.*, 2010) based on his/her interaction with this resource. In Mezghani *et al.* (2014), temperature is calculated over a period of time based on the user's annotation behavior for a resource. This temperature is related to three main parameters: freshness, which is relative to the dates of the annotated tags, popularity, which expresses the number of annotated tags and the similarity of users, which considers users who annotated the same resource.

In our work, we consider that friends have similar interests. From this perspective, we are basically interested in redefining the first-two parameters that are calculated for each interest ($i_a$) in $p_k$ ($p_k(i_a)$) belonging to the set of profiles ($\mathcal{P}$) across two equations. In fact, each of these two parameters expresses differently the importance of an interest. These parameters are introduced separately whether for the organization mechanism or the reliability mechanism.

The freshness parameter of an interest $p_k(i_a)$ is calculated by Equation (4). We take into consideration the date of $p_k(i_a)$ and the period of time between the minimum date and the maximum date of interests in $p_k$ which are, respectively, defined by Date($p_k(i_{\min})$) and Date ($p_k(i_{\max})$). Therefore, the value of freshness is a normalized date that is transformed into a value between 0 and 1:

$$\text{Freshness}(p_k(i_a)) = \frac{\text{Date}(p_k(i_a)) - \text{Date}(p_k(i_{\min}))}{\text{Date}(p_k(i_{\max})) - \text{Date}(p_k(i_{\min}))}. \tag{4}$$

The popularity parameter of an interest is used to express the number of interests $i_a$ existing in all profiles of $\mathcal{P}$. In fact, an interest $i_a$ can exist in a subset of $\mathcal{P}$ defined by $\mathcal{P}^+$. Hence, the popularity of $i_a$ is calculated by Equation (5) according to the size of $\mathcal{P}^+$, which is a value between 0 and 1:

$$\text{Popularity}(i_a) = \frac{\text{size}(\mathcal{P}^+)}{\text{size}(\mathcal{P})}. \tag{5}$$

## 4. Organization mechanism

The organization mechanism is one of the aspects of source reliability methods that corresponds to input data relative to interests in a profile $p_k$ belonging to $\mathcal{P}$. It is based on two processes: clustering and hierarchization. These two processes are identified in the Hierarchical-based Semantic and Temperature $k$-means algorithm (HSTK-means).

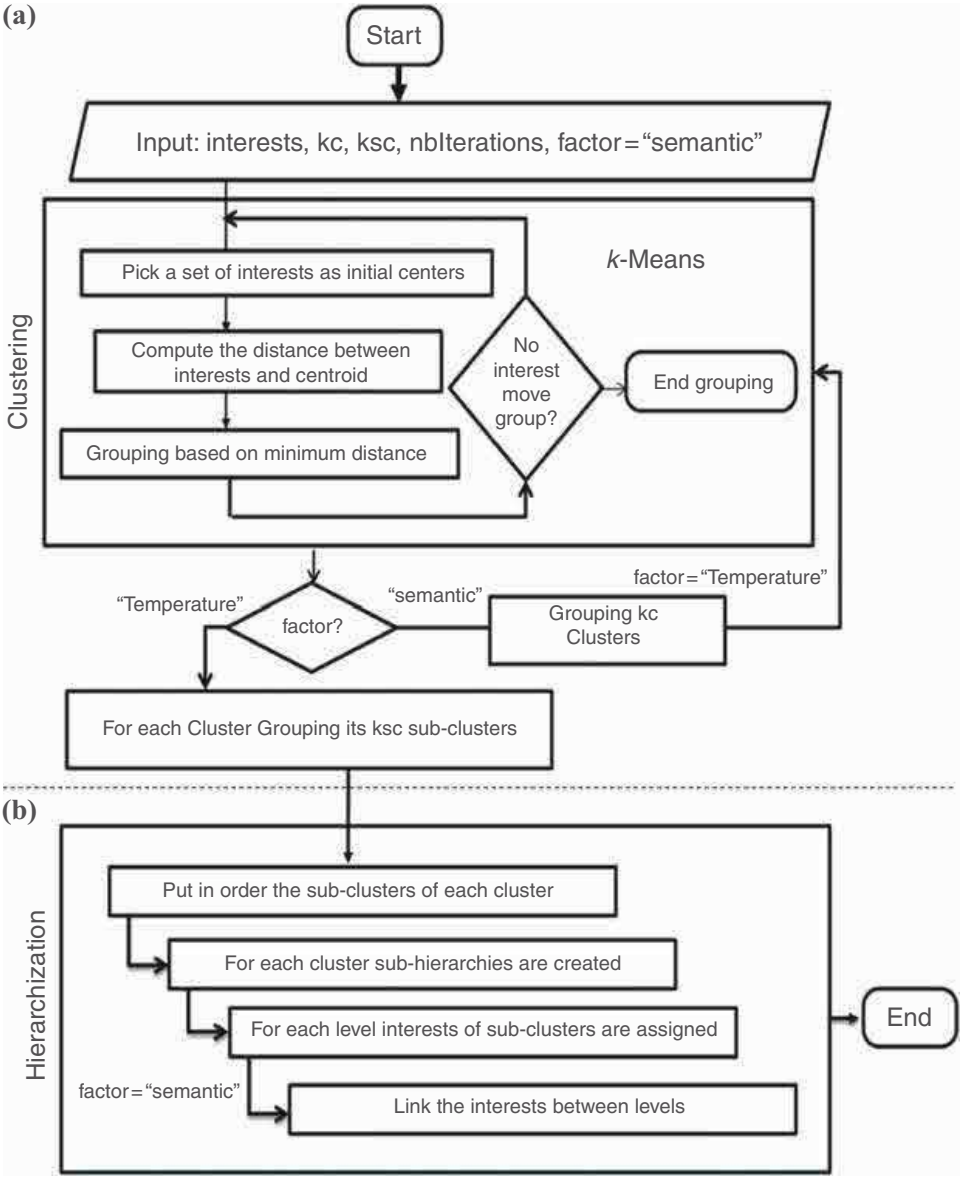
In order to better explain the organization mechanism, first, the clustering and the hierarchization processes are identified. Second, a scenario of their execution is presented.

### *4.1 Clustering process*

The clustering process allows to group the user's interests. It is illustrated in Figure 2(a). This process takes as input the interest set, the number of clusters "kc," the number of sub-clusters "ksc," the number of iterations "nbIteration," and the "factor" which is initialized by "semantic." It applies $k$-means to generate, on the one hand, for each user profile a set of n clusters $C = \{C_1, C_2, \ldots, C_n\}$, according to the semantic factor, in order to process the user's interests semantic link. In this case, $k$-means uses Equation (1) in order to compute the distance between interests (semantic distance). On the other hand, the clustering process applies the $k$-means algorithm to generate, for each $i$th cluster, a set of m sub-clusters $\text{SC}_i = \{\text{SC}_{i1}, \text{SC}_{i2}, \ldots, \text{SC}_{im}\}$, according to the temperature factor. Each sub-cluster contains a set of interests whose values of freshness as well as those of popularity are very close. In this case, $k$-means compute the distance between interests based on Euclidian distance by combining the freshness and popularity values that are presented, respectively, in Equations (4) and (5) (temperature distance).

### *4.2 Hierarchization process*

Generally, the representation of data in a hierarchical structure improves the result of source reliability in several works such as in Pasternack and Roth (2010) and Dong *et al.* (2014). However, the structuring is related only to the evolutionary characteristics of data and does not take into account the data semantic relationship. For this reason, the proposed hierarchization process allows the representation of the user's interests in not only a hierarchical structure but also in a semantic one based on clusters and sub-clusters generated by the clustering process. With this structure, the processing of interests becomes easier and more meaningful. The hierarchization process is illustrated




**Notes:** (a) Clustering process; (b) hierarchization process

**Figure 2.**
Organization mechanism (HSTk-means)

in Figure 2(b). It takes as input the result of the clustering process: the clusters ($C$) and their relative sub-clusters ($SC_i$).

First, sub-clusters $SC_i$ are ordered in descending order according to the values of freshness and popularity. Second, interests relative to sub-clusters are added into levels that depend on their scheduling. Finally, each interest existing in a sub-cluster is assigned to its direct parent based on the semantic factor (distance). The output of this step is a semantic tree that contains the hierarchy of interests constituted by sub-hierarchies (clusters).

### *4.3 Scenario of the organization mechanism (HSTK-means)*

We show the execution of the organization mechanism (HSTK-means algorithm) with a scenario instance illustrated in Figure 3. This figure is divided into four parts: (a), (b), (c) and (d).

Part (a) shows the input data constituted by a set of a user's interests. Each interest is characterized by a word ($W$), freshness ($F$) and popularity ($P$).

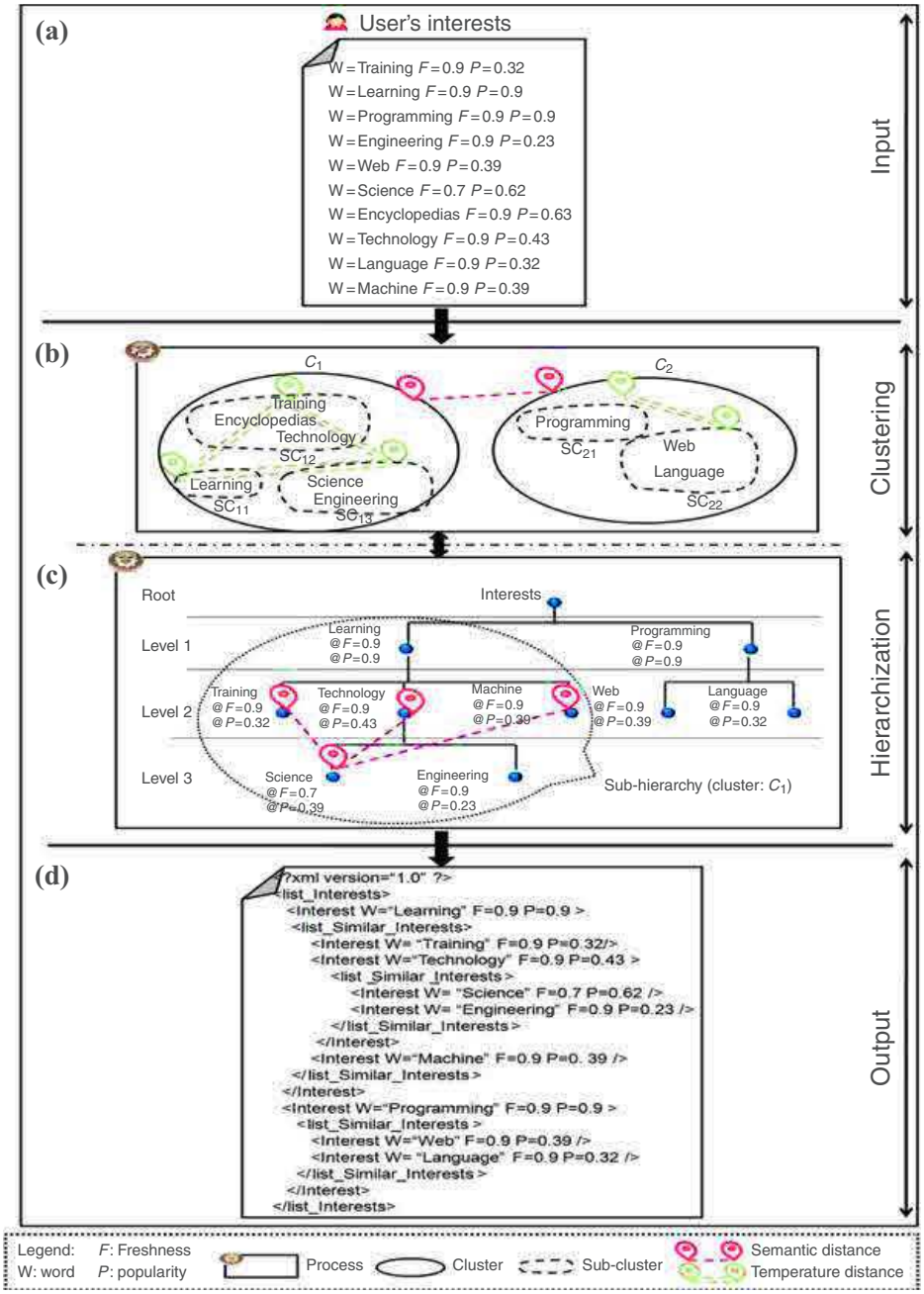



**Figure 3.**
An instance scenario of the organization mechanism

**Notes:** (a) Input; (b) clustering process; (c) hierarchization process; (d) output

In order to simplify the representation of the clustering process results, we illustrate, in part (b), the user's interests that are divided into only two clusters $C_1$ and $C_2$. For instance, the interests such as learning, science, machine, etc., which are semantically very close are included in the same cluster. Afterwards, $C_1$ ($C_2$) is divided into three (two) sub-clusters $SC_{11}$, $SC_{12}$ and $SC_{13}$ ($SC_{21}$ and $SC_{22}$). For example, the interests such as technology, machine and training, which have close popularity values and freshness values are included in the same sub-cluster $SC_{12}$.

Part (c) presents the hierarchization process which generates the hierarchy of interests according to clusters and their sub-clusters. Sub-clusters of each cluster are assigned to levels of the hierarchy while starting by adding the interests of the first sub-cluster.

For example, the first level of the hierarchy highlights the interests of the first sub-clusters $SC_{11}$ and $SC_{21}$. It contains the interest "Learning" and "Programming" which have the highest values of $F$ and $P$. The levels of the hierarchy are linked based on the semantic distance. For example, the interest "Science" existing in the third level is the closest semantically to "Technology" compared to "Machine" and "Training."

Part (d) presents how the interest hierarchy constituted by sub-hierarchies (clusters) is saved in an xml document. Each element (e.g. $<$ interest $W =$ "learning" $F = 0.9\ P = 0.9/>$) represents an interest and each interest is related to a list of similar interests ($<$ list_Similar_Interests $>$).

## 5. Reliability mechanism

The reliability mechanism is based on the last two aspects that are described in subsection 3.1. The first aspect is described by estimating the dependency between profiles through a profile dependency process. The second aspect predicts the reliability weight of each profile through a profile reliability weight prediction process.

For further clarification, the profile dependency and reliability weight prediction processes are identified followed by a scenario of their execution.

### *5.1 Profile dependency process*

The profile dependency process allows to quantify the copying relationship between two profiles $p_i$ and $p_j$ in the form of probability. The probability of dependency of $p_i$ on $p_j$ denoted $p(p_i \rightarrow p_j)$ is a weight that is calculated based on the profile dependency algorithm (cf. Figure 4(a)).

This algorithm differs from the state of the art algorithms in how to assess a data value (interest) that is copied from other sources and false. Moreover, the difference resides in the fact of using the semantic and hierarchical structure for the user's interests represented by their temperature factor.

The profile dependency algorithm takes into account two types of user's interest: local interest, which is not provided from the other profiles and distributed interest, which is provided from the other profiles. Distributed interest may be considered false during copying. In fact, a false value can propagate through copying, which can reduce the weight of reliability. However, a distributed interest may be considered true when its semantically corresponding interests have a better value of temperature.

This algorithm takes as input two organized profiles $p_i$ and $p_j$ and returns their probability of dependency. This probability is the quotient of the number of interests copied "nbCopiedInterests" by $p_i$ from $p_j$ by the total number of interests "nbTotalInterests" in $p_i$ (cf. below equation):

$$p(p_i \rightarrow p_j) = \frac{nb\text{CopiedInterests}(p_i)}{nb\text{TotalInterests}(p_j)}. \tag{6}$$

The number of copied interests (false interests) is calculated as follows: the algorithm checks the existence of each interest of $p_i(i_a)$ in $p_j$, in case of existence, $p_i(i_a)$ is considered as copied only if it is distributed in the other profiles and its freshness value in $p_i$ is lower than its freshness value in $p_j$, if $p_i(i_a)$ is a copied interest and has descendants in the hierarchy then the number of the copied interests is incremented by the total number of the descendants of $p_i(i_a)$. One possible reason for this incrementation could be related to the generated semantic and hierarchical structure of interests. In fact, an interest situated in an upper level has a freshness value higher than its descendants which are semantically very close. Thus, if an interest in an upper level is considered copied, it is reasonable that its descendants are also copied, the algorithm calculates the probability of dependency of $p_i$ on $p_j$ based on Equation (6).

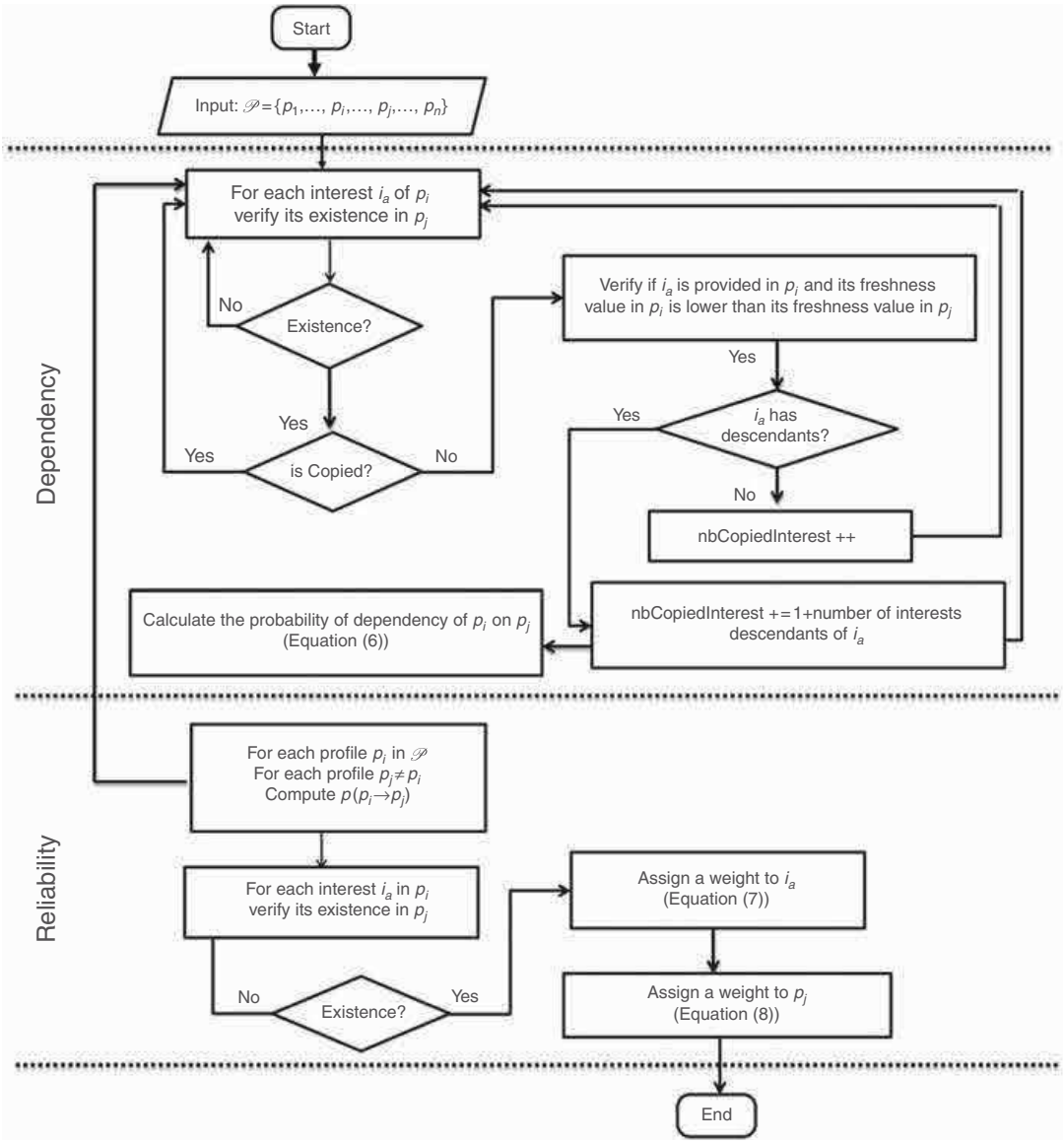



**Figure 4.**
Reliability mechanism

**Notes:** (a) Algorithm of the profile dependency process; (b) algorithm of the profile reliability process

### *5.2 Profile reliability weight prediction process*

The profile reliability weight prediction process identify the reliability weight of each profile in $\mathcal{P}$ based on the result of the profile dependency process. The algorithm of the profile reliability weight prediction process browses the semantic and hierarchical structure of interests relative to each profile $p_i$. Afterwards, it assigns a weight to each interest in $p_i$ by taking into account the temperature values and the result of the dependency between $p_i$ and each profile $p_j$ in $\mathcal{P}$ (cf. Equation (7)). Thus, the weight of each interest is calculated as the sum of the products of two main values relative to each profile in $\mathcal{P}$ (cf. Figure 4(b)).

The first value corresponds to the product of the freshness and popularity of an interest in $p_j$. The second value is the probability of non-dependency ($1-p(p_i \rightarrow p_j)$) which is proposed by source reliability approaches that assume the dependency between sources. Through this probability, the weight of dependency is subtracted from the first value (product of the

freshness and popularity). Therefore, the weight of the interest decreases which implies also the decrease of the reliability weight:

$$\text{Weight}(p_i(i_a)) = \frac{\sum_{j=0}^{\text{size}(\mathcal{P})} \text{Freshness}\left(p_j(i_a)\right) \times \text{Popularity}\left(p_j(i_a)\right) \times \left(1-p\left(p_i \rightarrow p_j\right)\right)}{\text{size}(\mathcal{P})}. \quad (7)$$

Finally, the algorithm aggregates the weights of interests of $p_i$ in order to compute its reliability weight (cf. Equation (8)). This weight is the quotient of the sum of its interest weights by the total number of interests in $p_i$ denoted $N$:

$$\text{Reliability}(p_i) = \frac{\sum_{n=1}^{N} \text{Weight}(p_i(i_n))}{N}. \quad (8)$$

*5.3 Scenario of reliability mechanism*

In this scenario, we consider three profiles in $\mathcal{P} = \{p_i, p_j, p_k\}$. We would compute the reliability weight of $p_i$. We assume that the probability of dependency, respectively, of $p_i$ on $p_j$ $p(p_i \rightarrow p_j)$ and $p_i$ on $p_k$ $(p(p_i \rightarrow p_k))$ are equal to "0.66" and "0.33." We also assume that the interest "Web" of $p_i$ is copied from $p_j$ and $p_k$ and its freshness and popularity values in $p_j$ and $p_k$, respectively, equal to "0.9" and "0.8." Therefore, the weight of the interest "Web" in $p_i$ is calculated based on Equation (7) as follows:

$$\begin{aligned}
\text{Weight}\left(\text{Web}\left[p_i\right]\right) &= \left(\text{Freshness}(\text{Web}\left[p_i\right]\right) \times \text{Popularity}\left(\text{Web}\left[p_i\right]\right) \\
&\quad + \text{Freshness}\left(\text{Web}\left[p_j\right]\right) \times \text{Popularity}\left(\text{Web}\left[p_{2j}\right]\right) \times \left(1-p\left(p_i \rightarrow p_j\right)\right) \\
&= (0.9 \times 0.8 + 0.9 \times 0.8 \times (1-0.66) + 0.9 \times 0.8 \times (1-0.33))/3 \\
&= 0.504
\end{aligned}$$

We notice that the weight of "Web" decreases from its initial weight ($0.9 \times 0.8$) in $p_i$ based on the probability of non-dependency. The reliability weight of $p_i$ is calculated by aggregating the weights of all the interests in $p_i$ based on Equation (8).

## 6. Evaluation

In this section, data sets and metrics used for the evaluation are described. Then, the obtained results are displayed.

*6.1 Data sets and metrics*

Our experiment was conducted on users who are at the same time members of the e-learning system Moodle (https://Moodle.org) and the social network Delicious (https://del.icio.us/).

Moodle contains the learner's interests which are explicitly provided by the learner or implicitly based on the visited and learned courses or lessons belonging to various domains. Delicious data set provides information about the user's friend relationships and the tagging behavior ($< user, tag, resource >$). A tag reflects a user's interest and may be related to an educational resource. Therefore, it enriches the user's interests in Moodle.

Regarding the evaluation, we extracted from Delicious the profiles of friends (explicit neighbors of each user) of a set of first-year university learners who belong to different sexes and who study in various areas, such as computer sciences, physics, etc. Each learner is provided with many links related to different pedagogic resources (courses, activities, etc.). Some of them do not match his/her current interests. For this reason, it is significant to recommend relevant pedagogic resources to the learner in order to support him while learning. In this social-learning context, recommendation results depend on learner's

interests that are enriched by the interests of his/her friends. These interests may be conflictual as they are coming from non-reliable profiles. In this evaluation, we assess the new profile reliability approach which we consider mandatory for each recommender system as it prevents the enrichment of interests from non-reliable profiles.

Figure 5(a) presents some characteristics of our data set.

The user's interests are represented in the form of matrix (cf. Figure 5(b)).

In each line, we find an interest which is described in three columns. The first column contains the values of interests represented by words. The second and the third columns contain, respectively, the freshness and popularity values.

We applied our proposed approach in order to detect the most reliable profiles so as to resolve conflicts between interests which can be enriched in the profile of a learner in Moodle. As mentioned in the previous sections, our approach rests on two mechanisms: organization (cf. Section 4) and reliability (cf. Section 5).

The first mechanism consists in creating a semantic and hierarchical structure of the user's interests based on two processes (clustering and hierarchization). The originality of this mechanism resides in merging temperature and semantic factors in HSTK-means algorithm.

Thus, two well-known evaluation metrics to assess the clustering result are selected: the Silhouette coefficient (Kaufman and Rousseeuw, 2009) and the Dunn Index (Dunn, 1974):

$$S(i\text{th}) = \frac{(b_i - a_i)}{\max(b_i; a_i)}. \tag{9}$$

The silhouette coefficient (Kaufman and Rousseeuw, 2009) allows to know to what extent an interest ($i$) belongs to its cluster. The silhouette coefficient is calculated for each interest in a cluster based on Equation (9) resting in turn on two values. The first value is relative to the average distance between each $i$th interest in a cluster and all other interests. This value is denoted $a_i$. The second value is relative to the average distance between the $i$th interest of a cluster and all clusters that do not contain this interest. This value is denoted $b_i$. The average of the silhouette coefficient of a cluster is calculated by taking the average of the silhouette coefficients of interests belonging to this cluster. A global measure of clustering relevance can be obtained by calculating the average of the silhouette coefficient of all clusters. A larger value means better clustering result.

**(a)**

| Description | Number |
|---|---|
| Number of users | 100 |
| Number of generated hierarchies (including users and their neighbors) | 545 |
| Average number of neighbors (profiles) per learner | 7 |
| Average number of interests per learner | 79 |

**(b)**

| | A | B | C |
|---|---|---|---|
| 1 | Interests (W) | Freshness (F) | Popularity (P) |
| 2 | Training | 0.9 | 0.32 |
| 3 | Learning | 0.9 | 0.9 |
| 4 | Novel | 0.8 | 0.46 |
| 5 | Lexicography | 0.3 | 0.20 |
| 6 | Programming | 0.9 | 0.9 |
| 7 | dictionary | 0.9 | 0.6 |
| 8 | Engineering | 0.9 | 0.23 |
| 9 | Web | 0.9 | 0.39 |
| 10 | Science | 0.7 | 0.62 |
| 11 | Encyclopedism | 0.9 | 0.63 |
| 12 | Technology | 0.9 | 0.43 |
| 13 | Language | 0.9 | 0.32 |
| 14 | Books | 0.9 | 0.23 |
| 15 | Article | 0.9 | 0.49 |

**Figure 5.**
Description of the data set (a) and an example of a user's interests (b)

The Dunn Index (Dunn, 1974) is introduced in order to recognize the well-separated and dense cluster. Let us denote by $d_{\min}$ the minimal distance between the interests of different clusters and $d_{\max}$ the largest distance within clusters. The Dunn Index is the ratio of $d_{\min}$ to $d_{\max}$ (cf. see below equation):

$$D = \frac{d_{\min}}{d_{\max}}. \tag{10}$$

If a data set contains well-separated clusters, the distance between the clusters $d_{\min}$ is generally large and $d_{\max}$ of the clusters are expected to be small. Therefore, a larger value means better clustering result.

After the clustering and sub-clustering step, HSTK-means algorithm creates the user's interest hierarchy which is in turn composed of sub-hierarchies that are evaluated on the basis of human judgment.

The second mechanism (cf. Section 5) consists in detecting the reliability weight for each profile of a learner's friend based on the profile dependency and profile reliability weight prediction processes. These processes rest on the semantic and hierarchical structure of interests. In order to evaluate the accuracy of the detected weights, we calculate the mean absolute error (MAE) and the root mean square error (RMSE) (Bobadilla *et al.*, 2013). To calculate these metrics, each user in Moodle is provided with the profile of his/her friends in Delicious for manual ranking. Afterwards, these profiles are ranked based on their reliability weights. MAE (cf. Equation (11)) is computed with the deviation between predicted rank ($p_i$) and manual rank ($r_i$) which constitutes sound truth of a profile:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |p_i - r_i|. \tag{11}$$

RMSE (cf. Equation (12)) is similar to MAE. What differs is that much more emphasis is put on larger deviation. Smaller MAE or RMSE indicates better accuracy:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (p_i - r_i)^2}. \tag{12}$$

After evaluating the organization and reliability mechanism, we evaluate the impact of the generated hierarchies on the reliability mechanism in terms of response time. The latter includes the time for computing the dependency between profiles and the profile reliability weights.

### *6.2 Evaluation of the organization mechanism*

In this section, we demonstrate the evaluation results relative to a sample of 20 users which are selected randomly. This evaluation consists of two phases.

The first phase aims at assessing the effectiveness of the semantic and temperature factors in terms of the clustering process.

The average of the silhouette coefficient values is measured for each user's generated clusters based on the semantic similarity factor. The semantic factor is calculated through the average of the IC and path length similarity measures (cf. Equation (1)). In other terms, we take the case where $\alpha = \beta = 0, 5$. These values are illustrated in Figure 6(a). The results indicate an improvement, for the majority of users, in the silhouette coefficient values based on the semantic factor compared to the results based only on the IC similarity measure ($\alpha = 1$ and $\beta = 0$). Noting that for some users (user 8, 17, 19) the IC similarity measure

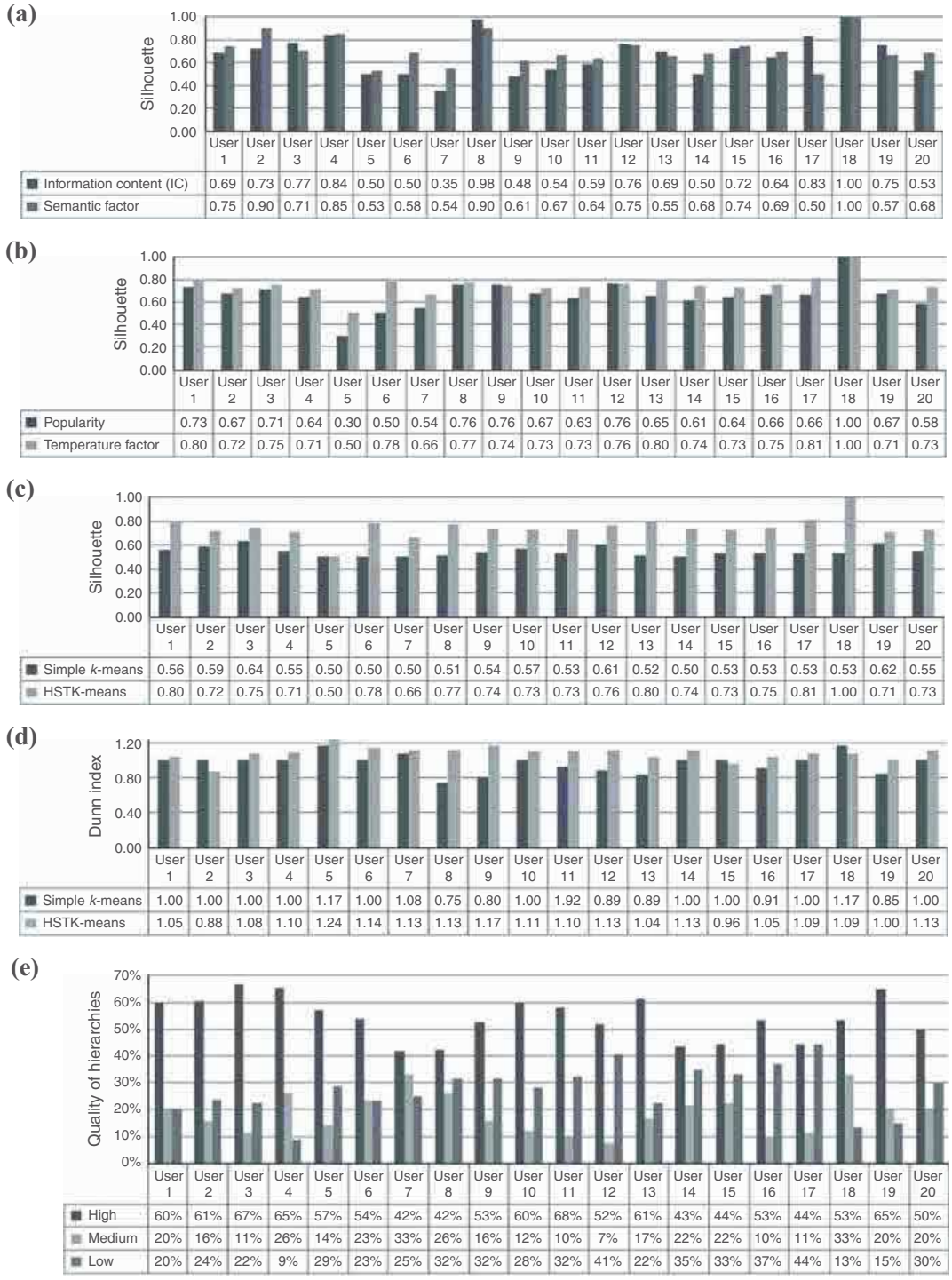



**Figure 6.**
Silhouette coefficient comparison values

**Notes:** (a) Between IC and semantic factor; (b) between popularity and temperature factor; (c) between simple *k*-means and HSTK-means; (d) dunn index comparison between simple *k*-means and HSTK-means; (e) the users' interest hierarchy quality levels

provides efficient silhouette coefficient values compared to the semantic factor, we found that the average of the silhouette values, with the semantic factor, rises from 67 to 70 percent. In order to improve this average, we need to adjust the parameters $\alpha$ and $\beta$ to highlight either IC or PL similarity measure.

Moreover, the average of the silhouette coefficient values is measured for each user's generated sub-clusters based on popularity and temperature factors. Figure 6(b) displays the generated results which exhibit a significant improvement. In fact, the average of the silhouette values for all users increases from 0.60 with popularity and reaches 0.75 with temperature factors.

Furthermore, the average of the silhouette and the Dunn Index values are measured for each user based on the simple *k*-means (which uses Euclidean distance as a factor) and our proposed HSTK-means algorithm (semantic and temperature factors). Figure 6(c) depicts a clear improvement in the silhouette values for all users since the average increases from 54 to 75 percent. In addition, the Dunn Index values show a clear improvement in Figure 6(d). We record an average of 0.97 for the simple *k*-means and 1.09 for our proposed HSTK-means. These results demonstrate that taking into account the semantic factor in addition to the temperature factor participates largely in the improvement of the generated clusters and sub-clusters.

The second phase aims at validating the created hierarchies for each user on the basis of human judgment. Figure 6(e) shows the evaluation result related to each user. There are three levels of hierarchy quality: high, medium and low. For each level, the percentage is specified with respect to the total created sub-hierarchies for each user.

For example, for user 1 there are 60 percent of sub-hierarchies with high quality, 20 percent are with medium quality and 20 percent are with low quality. Based on these evaluation values, we infer the percentage average value, for each level, relative to all the users' interest created hierarchies. We recorded that 54 percent of the total users' interests created hierarchies have high quality, 18 percent have medium quality and 28 percent have low quality.

The generated results (75 percent silhouette value, 1.09. Dunn Index and 54 percent hierarchies with high quality) prove that our organization mechanism is efficient.

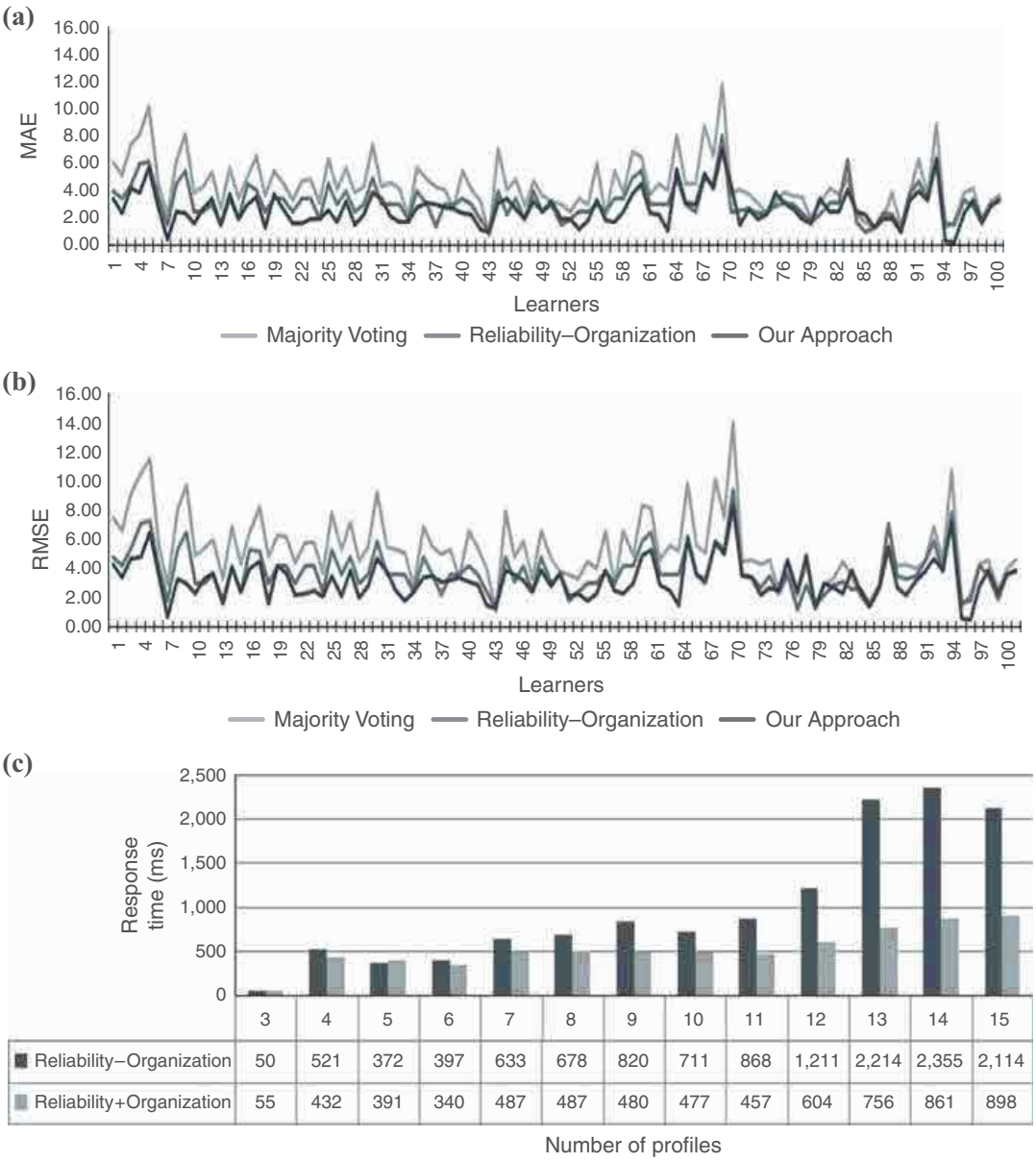
### *6.3 Evaluation of the reliability mechanism*

In this section, we demonstrate the evaluation results relative to 100 learners. Through this evaluation step, we propose to validate the importance of considering the semantic and hierarchical structure of interests in profile reliability weight detection. For this reason, we verify the potential and superiority of our approach compared to some existing approaches such as (Dong *et al.*, 2013; Huang and Wang, 2016; Li *et al.*, 2017) that do not take the semantic and hierarchical structure of data. For each user, we compute three values of MAE and RMSE. At a first stage, we evaluate the accuracy of the majority voting method which assigns highest weight to the profiles having highest number of interests that figure in the majority of other profiles. At a second stage, we assess the accuracy of reliability results without organization. Finally, we evaluate our approach which combines reliability and organization.

The results illustrated in Figure 7(a) and (b) portray a clear improvement of MAE and RMSE in our approach for all learners. For further clarification, MAE and RMSE values corresponding to a sample of ten learners are illustrated in a Table II.

We notice that results from the majority voting (average of 4.38) as well as those from reliability-based approach without organization (3.17) are not very satisfactory. One possible reason to account for these results is that in majority voting and previous reliability-based approaches, interests are regarded as separate (there is no relationship between interests whether in semantic or in temperature). However, the results obtained by considering the semantic interest hierarchy have clearly improved and are always the best with the lowest MAE average value (2.63).

The results of RMSE (cf. Figure 7(b)) confirm also that our approach remains the best with the lowest RMSE average value (3.15). This value indicates that there is a little deviation between the detected ranks and the manual ranks compared to other approaches in which we record an RMSE average values of 5.24 and 3.75.




| | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Reliability–Organization | 50 | 521 | 372 | 397 | 633 | 678 | 820 | 711 | 868 | 1,211 | 2,214 | 2,355 | 2,114 |
| Reliability+Organization | 55 | 432 | 391 | 340 | 487 | 487 | 480 | 477 | 457 | 604 | 756 | 861 | 898 |

**Figure 7.**
Profile reliability evaluations

**Notes:** (a, b) With MAE and RMSE; (c) with response time of profile reliability weight computing

After evaluating the MAE and RMSE values, we extracted the response time taken by the reliability mechanism for computing the reliability weights of the profiles of each learner's neighbors. Figure 7(c) illustrates the response time according to the number of profiles of neighbors. A clear improvement of response time where the reliability mechanism rests on the organization mechanism (semantic and hierarchical structure of the profiles). In fact, the average of response time decreases from 996 ms without organization to 518 ms with organization. Thus, we record 448 ms gain in response time which corresponds to 44.97 percent of time reduction.

These results imply that the use of the semantic interest hierarchy improves not only the reliability results (reduction of MAE and RMSE values according to previous approaches) but also the time taken to provide the reliability weights of profiles.

| Learners | MAE | | | RMSE | | |
|---|---|---|---|---|---|---|
| | MV | Rel-Org | Our approach | MV | Rel-Org | Our approach |
| 1 | 6.17 | 4.00 | 3.44 | 7.50 | 4.76 | 4.33 |
| 2 | 5.13 | 3.50 | 2.38 | 6.55 | 4.18 | 3.37 |
| 3 | 7.41 | 4.45 | 4.18 | 9.11 | 5.44 | 4.73 |
| 4 | 8.12 | 6.00 | 3.85 | 10.43 | 7.07 | 4.84 |
| 5 | 4.20 | 3.11 | 2.44 | 4.52 | 3.09 | 2.33 |
| 6 | 4.38 | 3.50 | 2.75 | 5.66 | 4.18 | 3.61 |
| 7 | 2.25 | 1.50 | 0.50 | 2.92 | 1.87 | 0.71 |
| 8 | 6.20 | 4.50 | 2.50 | 7.95 | 5.34 | 3.27 |
| 9 | 8.17 | 5.50 | 2.42 | 9.71 | 6.49 | 2.96 |
| 10 | 3.83 | 2.50 | 1.67 | 4.83 | 3.03 | 2.31 |

**Notes:** MV, majority voting; Rel-Org, reliability–organization; our approach, reliability+organization

**Table II.** Validation of the reliability mechanism with MAE and RMSE values

## 7. Conclusion

In this paper, we have detailed the proposed profile reliability approach attempting to resolve conflicts which may occur in neighbors' profiles of a user. This approach is based on two mechanisms which will be included in a recommender system: the organization mechanism which generates a semantic and hierarchical structure of each user's interests resting on two processes that are included in the hierarchical-based temperature and semantic $k$-means algorithm (HSTK-means), and the reliability mechanism that predicts the profile reliability weights. This weight is calculated grounded on the dependency weight between profiles which relies on the semantic and hierarchical structure of users' interests resulting from the organization mechanism.

We have experimented our approach, in social-learning context, based on learners/users in Moodle and Delicious. The generated results show the effectiveness of the proposed approach and prove that it can resolve conflicts among interests by predicting the reliability of profiles. For this reason, we consider that our approach is mandatory for each recommender system. Nevertheless, the proposed approach should be more improved. We should start by improving the semantic factor which needs the adjustment of parameters $\alpha$ and $\beta$. Besides, we should focus on other factors in addition to semantic and temperature.

On that account, in future works, we aspire to ensure the opening of e-learning systems and social networks on other systems, such as e-recruitment systems. These systems need the recommendation of the appropriate candidates for working. These candidates are recommended based on their learning experiences and their interests that are, respectively, stored in their distributed profiles and their curriculum vitae (CVs). As a consequence, a large number of profiles and CVs can be taken into account for conflict resolution. This may disturb the satisfaction of recruiters for such a candidate since many of his/her profiles and CVs may contain conflicting data. For this reason, we tend to consider, in addition to determining the reliability of profiles and CVs, two other concepts: intrusion detection (Peng *et al.*, 2016; Washha *et al.*, 2017) and serendipity (Kotkov *et al.*, 2016).

Intrusion detection systems are important for detecting and reacting to the presence of unauthorized users of a network or a system (Peng *et al.*, 2016). An intrusion detection system exploits the users' behavior (keystroke speeds, mouse use, language, preferences, etc.) in order to confirm or deny the legitimacy of their presence in the system. We suggest detecting unauthorized profiles and CVs as a first step before applying our approach.

Serendipity is a property that reflects how good a recommender system is at suggesting serendipitous items that are relevant, novel and unexpected. Novelty and unexpectedness require serendipitous items to be relatively unpopular and significantly different from a

user's profile (Kotkov *et al.*, 2016). For this reason, we suggest using this concept in the organization mechanism as well as the reliability mechanism.

In addition, as the number of profiles and CVs is evolving, we plan to implement the platform "Hadoop (http://hadoop.apache.org/)" which allows to analyze, store and manipulate large data (called big data).

## References


Barforoush, A.A., Shirazi, H. and Emami, H. (2017), "A new classification framework to evaluate the entity profiling on the web: past, present and future", *ACM Computing Surveys (CSUR)*, Vol. 50 No. 3, pp. 1-39.

Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A. (2013), "Recommender systems survey", *Knowledge-Based Systems*, Vol. 46 No. 10, pp. 109-132.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S. and Zhang, W. (2014), "Knowledge vault: a web-scale approach to probabilistic knowledge fusion", *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 601-610.

Dong, X.L., Berti-Equille, L. and Srivastava, D. (2013), "Data fusion: resolving conflicts from multiple sources", *Handbook of Data Quality*, Springer, Berlin and Heidelberg, pp. 293-318.

Dunn, J.C. (1974), "Well-separated clusters and optimal fuzzy partitions", *Journal of Cybernetics*, Vol. 4 No. 1, pp. 95-104.

Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F., Reiterer, S. and Stettinger, M. (2014), "Basic approaches in recommendation systems", *Recommendation Systems in Software Engineering*, Springer, Berlin and Heidelberg, pp. 15-37.

Gemmell, J., Shepitsen, A., Mobasher, B. and Burke, R. (2008), "Personalization in folksonomies based on tag clustering", *Intelligent Techniques for Web Person-Alization & Recommender Systems*, Vol. 12, pp. 37-48.

Ghorbel, L., Zayani, C.A. and Amous, I. (2016), "A novel architecture for learner's profiles interoperability", *Computer and Information Science 2015*, Vol. 614, Springer, Las Vegas, NV, pp. 97-108.

Gomaa, H.W. and Fahmy, A.A.A. (2013), "Survey of text similarity approaches", *International Journal of Computer Applications*, Vol. 68 No. 13, pp. 13-18.

Huang, C. and Wang, D. (2016), "Topic-aware social sensing with arbitrary source dependency graphs", *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*, *IEEE Press*, p. 7.

Huang, C., Wang, D. and Mann, B. (2017), "Towards social-aware interesting place finding in social sensing applications", *Knowledge-Based Systems*, Vol. 123, pp. 31-40.

Jadhav, M.R.R. and Wankhade, N. (2016), "A survey on recommender system", *International Journal for Research in Engineering Application and Management*, Vol. 2 No. 9, pp. 1-5.

Kalaï, A., Zayani, C.A., Amous, I. and Abdelghani, W. (2017), "Social collaborative service recommendation approach based on users trust and domainspecific expertise", *Future Generation Computer Systems*, Vol 80 No. 3, pp. 355-367.

Kaufman, L. and Rousseeuw, P.J. (2009), *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley Series in Probability and Statistics), Vol. 344, John Wiley & Sons.

Kotkov, D., Wang, S. and Veijalainen, J. (2016), "A survey of serendipity in recommender systems", *Knowledge-Based Systems*, Vol. 111 No. C, pp. 180-192.

Kumar, N., Rasool, A. and Hajela, G. (2016), "Keyword-aware hotel recommendation system", *International Journal of Computer Science and Information Security*, Vol. 14 No. 8, pp. 594-612.

Li, F., Lee, M.L. and Hsu, W. (2017), "Profiling entities over time in the presence of unreliable sources", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29 No. 7, pp. 1522-1535.

Li, H.-z., Hu, X.-g., Lin, Y.-j., He, W. and Pan, J.-h. (2016), "A social tag clustering method based on common co-occurrence group similarity", *Frontiers of Information Technology & Electronic Engineering*, Vol. 17 No. 2, pp. 122-134.

Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W. and Han, J. (2014), "Resolving conicts in heterogeneous data by truth discovery and source reliability estimation", *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 1187-1198.

Manzat, A., Grigoras, R. and Sèdes, F. (2010), "Towards a user-aware enrichment of multimedia metadata", Workshop on Semantic Multimedia Database Tech-nologies, pp. 30-41.

Martinez, M.L., Gonz_alez-Mendoza, M. and Valle, I.D.D. (2014), "Enrichment of learner profile with ubiquitous user model interoperability", *Computación y Sistemas*, Vol. 18 No. 2, pp. 359-374.

Mezghani, M., Péninou, A., Zayani, C.A., Amous, I. and Sèdes, F. (2014), "Dynamic enrichment of social users' interests", *IEEE 8th International Conference on Research Challenges in Information Science*, pp. 1-11.

Mezghani, M., Péninou, A., Zayani, C.A., Amous, I. and Sèdes, F. (2017), "Producing relevant interests from social networks by mining users' tagging behaviour: a first step towards adapting social information", *Data & Knowledge Engineering*, Vol. 108 No. 1, pp. 15-29.

Pasternack, J. and Roth, D. (2010), "Knowing what to believe (when you already know something)", *Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics*, pp. 877-885.

Peng, J., Choo, K.-K.R. and Ashman, H. (2016), "User profiling in intrusion detection: a review", *Journal of Network and Computer Applications*, Vol. 72 No. 14, pp. 14-27.

Rekatsinas, T., Dong, X.L. and Srivastava, D. (2014), "Characterizing and selecting fresh data sources", *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 919-930.

Rekatsinas, T., Joglekar, M., Garcia-Molina, H., Parameswaran, A. and Ré, C. (2017), "Slimfast: guaranteed results for data fusion and source reliability", *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1399-1414.

Sarma, A.D., Dong, X.L. and Halevy, A. (2011), "Data integration with dependent sources", *Proceedings of the 14th International Conference on Extending Database Technology*, pp. 401-412.

Simpson, E. (2008), "Clustering tags in enterprise and web folksonomies", *Proceedings of the Second International Conference on Weblogs and Social Media*, pp. 222-223.

Varma, S., Sameer, N. and Chowdary, C.R. (2017), "Relic: entity profiling by using random forest and trustworthiness of a source-technical report".

Walsh, E., O'Connor, A. and Wade, V. (2013), "The fuse domain-aware approach to user model interoperability: a comparative study", *IEEE 14th International Conference on Information Reuse and Integration*, pp. 554-561.

Washha, M., Qaroush, A., Mezghani, M. and Sèdes, F. (2017), "A topic-based hidden Markov model for real-time spam tweets filtering (regular paper)", *International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, *Marseille*, *Science Direct*, *September 6–8*, pp. 833-843, available at: www.sciencedirect.com

Xu, G.D., Zong, Y. and Jin, P. (2015), "KIPTC: a kernel information propagation tag clustering algorithm", *Journal of Intelligent Information Systems*, Vol. 45 No. 1, pp. 95-112.

Yin, X. and Tan, W. (2011), "Semi-supervised truth discovery", *Proceedings of the 20th International Conference on World Wide Web*, pp. 217-226.

Yu, D., Huang, H., Cassidy, T., Ji, H., Wang, C., Zhi, S., Han, J., Voss, C. and Magdon-Ismail, M. (2014), "The wisdom of minority: unsupervised slot filling validation based on multi-dimensional truth-finding", *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1567-1578.

Yu, H., Zhou, B., Deng, M. and Hu, F. (2018), "Tag recommendation method in folksonomy based on user tagging status", *Journal of Intelligent Information Systems*, Vol. 50 No. 3, pp. 479 -500.

Zhang, H., Li, Q., Ma, F., Xiao, H., Li, Y., Gao, J. and Su, L. (2016), "Inuenceaware truth discovery", *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 851-860.

Zhao, Z., Cheng, J. and Ng, W. (2014), "Truth discovery in data streams: a singlepass probabilistic approach", *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pp. 1589-1598.

**Corresponding author**
Leila Ghorbel can be contacted at: leila.ghorbel@gmail.com