I.J. Information Technology and Computer Science, **2014, 12, 35-45**
Published Online November 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2014.12.05

An Integrated Approach to Drive Ontological Structure from Folksonomie

Zahia Marouf, Sidi Mohamed Benslimane
EEDIS Laboratory, Djillali Liabes University of Sidi Bel Abbes, Sidi Bel Abbes, 22000, Algeria
Email: {marouf.zahia, benslimane }@univ-sba.dz

Abstract— Web 2.0 is an evolution toward a more social, interactive and collaborative web, where user is at the center of service in terms of publications and reactions. This transforms the user from his old status as a consumer to a new one as a producer. Folksonomies are one of the technologies of Web 2.0 that permit users to annotate resources on the Web. This is done by allowing users to use any keyword or tag that they find relevant. Although folksonomies require a context-independent and inter-subjective definition of meaning, many researchers have proven the existence of an implicit semantics in these unstructured data. In this paper, we propose an improvement of our previous approach to extract ontological structures from folksonomies. The major contributions of this paper are a Normalized Co-occurrences in Distinct Users (NCDU) similarity measure, and a new algorithm to define context of tags and detect ambiguous ones. We compared our similarity measure to a widely used method for identifying similar tags based on the cosine measure. We also compared the new algorithm with the Fuzzy Clustering Algorithm (FCM) used in our original approach. The evaluation shows promising results and emphasizes the advantage of our approach.

Index Terms— Folksonomies, Collaborative Tagging, Ontologies, Fuzzy Clustering, Similarity Measure.

I. Introduction

Web 2.0 is the second generation of Internet based services that emphasizes the role of users on the Web. Users are encouraged to add content, manage it, and share it with other users in an interactive and collaborative way. This blurs the boundaries between Web users and producers, consumption and participation, authority and amateurism, play and work, data and the network, reality and virtuality [1]. Web 2.0 applications are built around user-generated or user-manipulated content, such as wikis, blogs, podcasts, social networking sites, and collaborative tagging systems or folksonomies.

Folksonomies [2] have recently emerged as a powerful way to label and organize large collections of data. These systems allow users to use any keywords or tags relevant to the content to annotate their favorite resources on the web. When a certain number of users annotate an item with a tag, this starts to look like a reasonable description of the item, and forms a consensus about the tag. This means that folksonomies contain implicit evidences for the underling semantics that can be exploited as complement to more formalized Semantic Web technologies.

Folksonomies are created by users without contribution of experts. This makes them relatively easy and rapid to build, conceptua1lly simple, cheap, facile to use, and highly scalable. In despite of these advantages, folksonomies include all kinds of tags varying from standard dictionary words and compound expressions to jargon and nonsense words. As a result, they contain ambiguous, overly personalized and imprecise words.

Various solutions have been proposed to make the emergent semantics in folksonomies explicit [3, 4, 5, 6, 7, 8]. Clustering approaches are widely used but most of them use co-occurrence count either with users or with resources, which means that they do not deal with the three modes of folksonomies. Furthermore, these approaches do not give a formal solution to the disambiguation and context identification problem. Another stream of research associates semantic entities to tags as a way to formally define their meaning, but these approaches need existing ontologies that match well the folksonomy

In this paper, we propose an improvement of our approach for extracting hierarchies from folksonomies previously introduced in [9]. An ameliorated similarity measure as well as a new algorithm for context identification and disambiguation are introduced. These contributions emphasize the advantage of our novel approach by overcoming the limitations of the original as well of other approaches.

The rest of the paper is structured as follows. In section 2, we provide an overview of related work on acquisition of semantics from folksonomies. In section 3, we outline the proposed approach and discuss the detailed steps. Section 4 introduces an experimental methodology to evaluate the approach. Finally, section 5 concludes the paper and points directions for future work.

II. Related Work

The origins of automatic acquisition of semantics from unstructured and semi-structured resources can be found in ontology learning from text [10]. Existing approaches to infer hierarchical tag relationships from folksonomies can broadly be assigned to one of the following classes:

A. Clustering approaches

These approaches identify the semantics of tags, by clustering tags according to some relations among them. Mika [3] describes an approach to generate lightweight ontologies from folksonomies based on the overlapping set of users, and the overlapping set of resources. Hamasaki et al [4] extended Mika's work by taking into account tagging information of the user neighbors in the folksonomy. Begelman et al.'s approach [5] proposes to split the co-occurrence graph in two clusters with a technique called Spectral bisection, and to execute this technique recursively on the new clusters. Kennedy et al [6] present a clustering algorithm applied on spatial and temporal distributions to find groups of tags sharing spatial or temporal patterns. Heyman et al [7] use Cosine similarity between tags to measure the distance from one tag to another, then organize them into a hierarchical tree by starting with a single "root" node, and adding other tags to the tree in decreasing order of centrality. Benz et al [8] have proposed an extension of this algorithm. The authors add context identification and disambiguation tasks to the algorithm. A promising approach is presented in [9]. The authors propose a new similarity measure called CDU (Co-occurrences in Distinct Users), that exploits the three mode of the folksonomy. After cleaning the data, tags are represented in a vector space to calculate the cosine matrix. A fuzzy clustering algorithm FCM [11] is performed on the cosine matrix for disambiguating and identification of context of tags. Context of a tag is the set of tags in the same cluster, while ambiguous tags are the ones belonging to the intersection of clusters. In the last step, they ameliorate the algorithm of Heymann et al [7] by using anew generality measure called FDU (Frequency by Distinct Users) to extract the hierarchy of tags. In section 3, we discuss the limitations of this approach.

These clustering approaches measure tags similarity based on the resource regardless of the annotator, or on the user regardless of the resource, and most of them do not deal with ambiguity problem or do not give a formal solution to it. Furthermore, most of them do not make the hierarchical relations explicit between tags.

B. Association rule mining approaches

This class of approaches applies association rule mining techniques by discovering knowledge that is already implicitly present.

Schmitz et al [12] propose a systematic overview of projecting a folksonomy onto a two-dimensional structure. Then they show the results of mining rules from selected projections. In [13], Jäschke et al extend the data-mining task of discovering all closed item sets to three-dimensional data structures. Their algorithm returns a tri-ordered set of triples called triadic concepts in Formal Concept Analysis FCA [14], where each triple consists of a set of users, a set of tags, and a set of resources. Trabelsi et al [15] introduce a new algorithm, called Tricons that directly tackles the triadic form of folksonomies towards a scalable extraction of tri-concepts.

These approaches output a hierarchical representation of tags, but the relationships between tags in different hierarchical levels are not defined semantically, and there is no strategy to deal with ambiguous tags.

C. Semantic based approaches

These approaches aim at associating semantic entities with tags as a way to formally define their meaning. Angeletou et al [16] propose an automatic approach to enrich folksonomy tags with formal semantics by associating them with relevant concepts defined in online ontologies. Cantador et al [17] present an automatic approach to associate folksonomy tags with domain ontology concepts using Wikipedia[1] categories. Garcia-Silva et al [18] proposed an approach to link tags to DBpedia [19] resources by means of selecting the Wikipedia page that best represents the tag. Djuana et al [20] present personalization strategies to disambiguate tags by combining the opinion of WordNe [2] lexicographers and users' tagging behavior together.

This class of approaches needs an existing upper ontology as the base structure. The lack of ontologies that well match the tags in folksonomies is one of the major obstacles applying these approaches.

D. Hybrid approaches

In this section, we present some approaches integrating multiple techniques. Giannakidou et al [21] cluster tags based on a similarity measure that mixes tag co-occurrence with semantic similarity extracted from ontologies. This approach clusters tags into disjoint groups. This means that for an ambiguous tag, the approach will only identify the most frequent meaning according to the tag co-occurrence pattern.

Specia and Motta [22] propose a semi-automatic approach that clusters tags based on the co-occurrence information, and maps them into ontology elements (concepts, properties, instances, etc.). However, the clustering task in this approach doesn't appear ambiguous tags, so a disambiguation process is executed to analyze each cluster and detect ambiguities. Moreover, the semantic identification activity in this approach is performed manually.

Lin et al [23] propose an approach that exploits the power of low support association rule mining supplemented by an upper ontology such as WordNet. However, the approach filters out non-English words in the preprocessing step, so it doesn't deal with the multilingual structure of the folksonomy.

Schmitz et al [24] presented a formal model of folksonomies as a set of triples or, equivalently, a tripartite hypergraph. In order to apply association rule mining to folksonomies, they have systematically explored possible projections of the folksonomy structure into the standard notion of "shopping baskets" used in rule mining. This approach is however limited as it doesn't present any disambiguation activity.

III. The Proposed Approach

The work presented in this paper relies on a novel combination of similar techniques. It deviates from other

[1] http://en.wikipedia.org/

[2] http://wordnet.princeton.edu/

approaches by using novel similarity and generality measures. Furthermore, it uses fuzzy clustering instead of the hard one to define context of tags and disambiguate ambiguous tags.

In this section, we describe an extension to our approach that aims to extract ontological structures from folksonomies. The new approach has the same steps as the old one, but it overcomes some limitations that are discussed one by one in the sections describing the different steps. The overall process is depicted in the Fig.1.

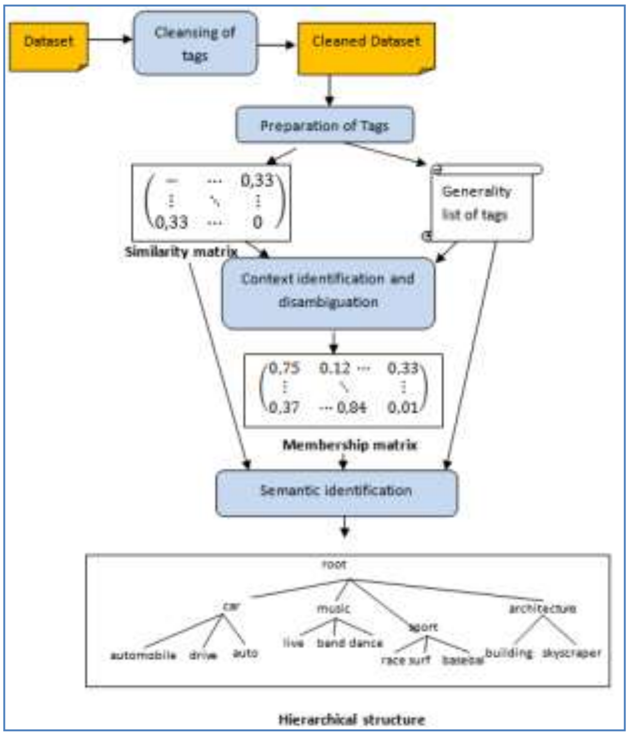

Fig. 1.The proposed approach

A. Cleansing of Tags

Since actors can choose any keyword for categorizing their content, they are applying their own spelling and tagging rules (e.g. singular or plural nouns, conjugated verbs). Consequently, tags are polluted and need to be cleansed. Therefore, before analyzing all the data sets of folksonomies, we must clean tag sets. For this purpose, we proceed as described in the algorithm 1.

In this paper, we deal only with English tags to have coherent hierarchical structure. A multilingual solution is to generate as hierarchical structures as languages count used in the folksonomy.

Algorithm Cleansing

```
Inputs: datasetFile
Outputs :cleanDataSetFile
while not end (datasetFile)  do
DatasetFile.readLine()
        if  tag is  in{stop words, meaningless tags, infrequent tags} then
DatasetFile.getNextline
        else
               Exist = Wordnet.check(tag);
        /* look for the tag in WordNet*/
                 if not Exist then
                  end if
tag =Stemming(tag)
/*The stemming task reduces tags to their stem or root*/
          end if
CleanDataSetFile.add(Tag)
end while
```

B. Preparation of tags

In this step, we generate a list of tags in their generality order and a vector space representation of tags. The generality degree is based on the FDU measure that counts the frequency of use of a given tag by distinct users. The vector space representation is based on an ameliorated version of our proposed similarity measure CDU (Co-occurrences in Distinct Users). Although this later outperforms the Co-occurrence measure and ameliorates results when used as the basis to calculate other similarity measures as the cosine one, it suffers from having non-normalized values, so that it cannot be the input for other steps such as clustering and hierarchy extraction. These steps are based on comparison between similarity values that require to be normalized. In the prior work, we tackled this problem by calculating the cosine similarity matrix using the CDU matrix. We call this new version of similarity measure NCDU (Normalized CDU).

Before describing the FDU and NCDU measures, we first give a definition of folksonomy and its elements.

Definition 1: Folksonomy

A folksonomy is defined as a tuple $F = \{T, U, R, Y\}$, where T is the set of tags that comprise the vocabulary expressed by the folksonomy; U, R are respectively the sets of users and resources that annotate and are annotated with the tags of T; and $Y = \{(u, t, r)\} \in U \times T \times R$ is the set of assignments (annotations) of each tag to a resource by a user u.

A post is a triple (u, t_{ur}, r) with $u \in U$, $r \in R$ and anon-empty set $t_{ur} = \{ t \in T \setminus (u, r, t) \in Y \}$.

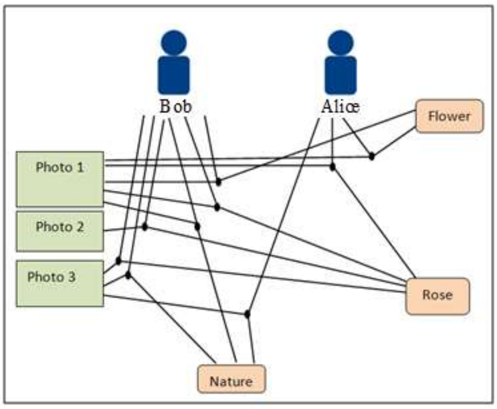

Fig. 2. An example of a folksonomy

Fig. 2 displays an example of a folksonomy. In the following, we use this running example to illustrate the proposed measure.

Generality measure: A first natural intuition is that the more general tags are simply the more often used ones, since they are well known by users. We capture this intuition in the generality measure “FDU”.

Definition 2: FDU generality measure

The FDU generality measure is an adapted version of frequency that counts number of distinct users annotating resources with a given tag. Formally, we define it as follows:

$$\forall t, \in T , u \in U, FDU(t) = card\{(u) \in U | t \in t_{ur}\} \quad (1)$$

Below, we report the resulting similarity FDU values of the example.

Table 1. The resulting FDU values

Tags	Flower	Rose	Nature
FDU	2	2	2

Similarity measure: As measures for relatedness are not well developed for three-mode data such as folksonomies, we need to narrow triples by one. Various solutions have been proposed to calculate similarities between tags and resources based on two-mode views of the data [25]. Unlike the prior work where we used the macro-aggregation solution, in this paper we use a tag-tag binary representation for each user instead of the tag-resource representation, then we aggregate across users. This representation ameliorates runtime performance, because the resources’ number is significantly bigger than tags one (especially after the cleansing task), so the tag–resource representation necessitates more space memory and calculating time.

The values of the tag-tag per-user binary representations are $w_u\,(t_1\,, t_2) \in \{0.1\,\}$, where t_1 and t_2 are pairs of tags.$w_u\,(t_1\,, t_2) = 0$ means that t1and t_2 do not appear together in any post associated to the user “u”.$w_u\,(t_1\,, t_2) = 1$ means that t_1 and t_2 co-occur at least once in the posts of the user. Formally*:*

$$\forall (t_1\,, t_2) \in T\,,\, u \in U,\, r \in R$$

$$w_u(t_1\,, t_2) = \begin{cases} 1 & if \exists\, t_{ur} : (t_1\,, t_2) \in t_{ur} \\ 0 & else \end{cases} \quad (2)$$

For the folksonomy example in Fig.2, the binary representation is shown in Table 2 and Table 3 for the users Bob and Alice respectively.

Table 2. Bob’s binary matrix

Tags Tags	Flower	Rose	Nature
Flower	-	1	1
Rose	1	-	1
Nature	1	1	-

Table 3. Alice’s binary matrix

Tags Tags	Tags	Rose	Nature
Flower	-	1	0
Rose	1	-	0
Nature	0	0	-

NCDU is an ameliorated version of CDU. It is calculated by summing the binary matrix across users, then dividing on the smallest generality degree of each pair of tags. The advantage of this second version of CDU is not restricted only on considering the three modes of a folksonomy all at once, but also on generating normalized and more accurate values of similarity.

Definition 3 : NCDU Similarity measure

The NCDU similarity measure is an adapted version of CDU. Formally, we define it as follows:

∀ t1, t2 ∈T , u ∈U, r ∈R:

$$NCDU(t_1\,, t_2) = \frac{\sum_u w_u(t_1\,, t_2)}{\min(FDU(t_1), FDU(t_2))} \quad (3)$$

Below, we report the resulting similarity matrix NCDU in table 4.

Table 4. The resulting NCDU matrix

Tags Tags	Flower	Rose	Nature
Flower	-	1	1/2
Rose	1	-	1/2
Nature	1/2	1/2	-

This new measure permits us to compare values easily in the step of generating the hierarchy. Moreover, it avoids calculating another matrix as cosine matrix or others. This is because of the normalization as well as the flexibility that offers our new algorithm of clustering which deals with spares matrices. This algorithm is the focus of the next section.

C. Context identification and disambiguation

Despite the advantages of social tags, they suffer from various vocabulary problems. Ambiguity (polysemy) of the tags arises as users apply the same tag in different domains. On the other side, the lack of synonym control can lead to different tags being used for the same concept. These problems are being investigated in the literature. There are approaches that attempt to identify the actual meaning of a tag by linking it with structured knowledge bases [22, 26].

Other works apply probabilistic models and clustering techniques on the tag space according to the tag co-occurrences in item annotation profiles [27, 28, 29].

In this paper, we follow a clustering strategy as well, but in contrast to previous approaches, our proposition provides the following benefits:

- Instead of using standard clustering processes, we propose to apply a fuzzy clustering technique that allows tags to belong to more than one cluster. Similar tags are belonging to the same cluster, whereas ambiguous tags are found in the intersection of two or more clusters.
- Instead of using simple tag co-occurrence similarity measure, we use the new similarity measure NCDU**.**

The literature provides many examples for context identification and disambiguation. These approaches consider hard clustering techniques as a way to assign a set of tags into distinct clusters so that tags in the same cluster are more similar to each other than to those in other clusters. This clustering task figures the meaning of tags by collecting his similar tags or his context, but supplement efforts must be perused to tackle ambiguity problem. In [22] the disambiguation activity analyzes each group of similar tags found in the context identification in order to find tags with different meanings. Benz et al [8] applied a standard average-link hierarchical algorithm [30] to disambiguate tags.

Inspired by prior work, and after having limited success-producing clusters with other algorithms, we developed Algorithm 2, a new fuzzy clustering algorithm to group similar tags into clusters and to identify ambiguous tags that belong to more than one cluster.

Algorithm 2: context identification and disambiguation of tags

```
Inputs:
- NCDU[n][n] /*The similarity matrix of tags $t_1,…,t_n$*/
- $L_{generality}$/*a list of tags $t_1,…,t_n$ in descending vertex
  order of their generality measured by FDU */
- min_sim/* parameter for the threshold at which a tag is
  chosen as a center */
Output:
- c /*  The number of clusters generated*/
- U[n][c] /* the membership matrix of the tags*/
1: centers [0] $\leftarrow L_{generality}[0]$ /* add the first tag in the
   generality list as the  first center */
2: c=1 /* the actual number of clusters*/
3:i=1;     /* index of $L_{generality}$*/
4: while i<$L_{generality}$.size do
   /* testing if the similarity between the new center and
   the othersis less than the similarity threshold to minimize
   the similaritybetween the generated clusters*/
5:   if NCDU[$L_{generality}$[i]][allcenters]<min_sim then
6:           c+ =1
7:           centers[c]$\leftarrow L_{generality}$[i]
8:   end if
9:   i+=1 /* take another tag from the $L_{generality}$*/
10:end while
11: for i=0 tot do
12:   for k=0 to c do
13:     U=NCDU[i][k] /* membership values*/
14:   end for
15:end for
16: return (c, U)
```

We first discuss the algorithm then we offer some insight why such a simple algorithm is extremely successful. The algorithm starts by adding the first tag in the generality list as the first center (line 1), and initializing the number of clusters "c" (line 2). After that, it chooses tags from the generality list to be considered as centers. This choice is based on a similarity threshold to ensure maximizing the distances between centers (lines 4-10). Once all centers are chosen, the algorithms calculates the membership matrix to be returned with the number of clusters as outputs.

This new algorithm overcomes the limitations of the Fuzzy C-means clustering (FCM) [11] and of most of clustering algorithms. We cite these limitations below:

- The number of clusters "c" must be predefined: the main disadvantage of FCM and of the majority of clustering algorithms is the obligation to define a fixed number of clusters. This task is very difficult whatever the nature of the data, and more difficult in our case because defining how many clusters can be generated when clustering a set of tags in a folksonomy is not a trivial task.
- The value of the fuzziness degree "m" in FCM is a prerequisite for the algorithm: in the case of the disambiguation task, this parameter represents how many senses an ambiguous tag can have in a given dataset. The value of this parameter is often 2, which means that no tag can have more than two different senses, and all ambiguous tags have the same meanings count.
- Attributing tags to clusters is based on the distance between these tags and the centers of clusters, but these centers are not effective tags, so the context of a tag is based on the other tags in the same cluster, so is not clearly defined because a cluster can have noisy tags.
- The results as sensitive to the initial guess, like the centers and the membership matrix, and a good choice for this guess is not evident.
- Clustering algorithms have mostly long computational time (complicated calculations, and lot of iterations).

Our algorithm overcomes all these shortcomings as it decides by itself the number of clusters. This is achieved by adding new centers while there are tags not classed. It also doesn't require any value of the fuzziness parameter, so it extracts the effective number of senses that an ambiguous tag has depending on the similarity values between a tag and centers of the associated clusters. In addition, the context extracted by our algorithm for the tags is clearly defined since centers of clusters are a set of chosen tags from the folksonomy, so they can be considered as the definitions of meanings of clusters. Furthermore, unlike the other algorithms of clustering, this new algorithm doesn't need any initial guess as it doesn't use any arbitrary information, so the results do not change when rerunning the algorithm except if the dataset is changed.

Moreover, this new algorithm is simple and doesn't necessitate long computational time. This will be demonstrated in the experimental section by comparing it with FCM. At the end of the context identification and disambiguation step, we have a set of overlapped clusters. The tags that belong to the intersection are ambiguous tags, and the number of their meanings is the associated clusters count. Once this step is achieved, we can generate the hierarchy of tags. This task is discussed in the next section.

D.Semantic identification

Most of the approaches associating semantic entities to tags rely on string matching techniques to find candidate ontology concepts and then use the tag context to choose the one that better describes the meaning of a tag. However, this activity implies the transition from a flat space, i.e., without hierarchies, in the folksonomy side, to a hierarchical space in the ontology side. Some works tackle this problem by associating tags initially to WordNet Synsets, and then the Synset hierarchical structure is compared against ontologies [16]. In a comparative study realized by [31], it was proven that the algorithm of Heymann et al [7] outperforms all the algorithms introduced in the study.

In this paper, we propose Algorithm 3 that extend Heymann’s algorithm by providing the following improvements:

- Instead of using generality measured by degree centrality in the tag-tag co-occurrence network [32] as in [8], we use FDU as generality measure. Thereby, we take the dimension of the user into account.
- Tags underneath the root are the centers generated in the context identification and disambiguation step.

Algorithm 3: Semantic Identification

```
Inputs:
- Lgenerality/*a list of tags $t_1,\dots,t_n$ in descending vertex
  order of their generality measured by FDU */
- nc[$t_1$],…,nc[$t_n$]/* The count of clusters nc[$t_1$],…, nc[$t_n$]
  where the tag is included */
- min_sim/* parameter for the threshold at which a tag
  becomes a child of a related parent rather than the root
  */
- c /* the centers count*/
- Centers[c] /* the centers set generated in the clustering
  step*/
Output:
- Hierarchy
1: Hierarchy←< root>
2: for i=1 to c do
3:    Hierarchy←centers[i]
4:Lgenerality. Remove(centers[i]);
/* remove the tag from the generality list*/
5: end for
6: for i=0 to |$L_{generality}$|do
7:     $t_i$ ← $L_{generality}$[i]
8:     k=nc[$t_i$]
9:     maxCandidateVal ← 0
10:    repeat /*if $t_i$ is an ambiguous tag, repeat steps 7 to
17 for each of its senses.*/
11:      for all $t_j \in$getVertices(Hierarchy) do
/*identify the most similar existing tag $t_i$ to $t_i$*/
12:        if CDUN($t_i$,$t_j$ ) >maxCandidateVal then
/* computes the cosine similarity between ti and tj */
13:            maxCandidateVal ←NCDU($t_i$,$t_j$ )
14:            maxCandidate ← $t_j$
15:        end if
16:end for
17:  if maxCandidateVal>min_sim then
/* $t_j$ is the most similar tag to $t_i$*/
18:         Hierarchy ← Hierarchy ∪<maxCandidate,$t_i$>
/* $t_i$ is added to the hierarchy under $t_j$*/
19:  else
20:         Hierarchy ← Hierarchy <root,$t_i$>
21:  endif
22:k=k-1
23:  Until k =0 /* all senses of $t_i$ are added to the
hierarchy*/
24: end for
25: Return (Hierarchy)
```

The algorithm starts with a tree with a single node “root” that represents the top of the tree (line1). Then the centers generated by the clustering algorithm are put underneath the root without forgetting to remove them from the generality list (lines 2-5). Each tag is then added in the decreasing order of how general the tag is. The algorithm adds each tag to his most similar one in the tree if their similarity is greater than a similarity threshold (lines 7-18), else it is added under the root (lines 19-20). If the tag is ambiguous then it repeats steps 10-23 for each of its senses.

IV. Experiments

We perform a set of experiments in order to quantify the influence of the proposed measures of tag relatedness, and tag generality on emergent tag semantics in a folksonomy. As well as to evaluate the ability of the new clustering algorithm to define context of tags and to detect ambiguous ones. We will first provide details on our dataset and then explain each experimentation step before discussing the results.

A. Dataset

We used a snapshot of the Flickr3 metadata database gathered in the period from November 24th to December 31th, 2005. Originally it contained |U|= 111, 920 users, |R|= 3,253,390 resources, and |T|= 374,076 tags. In the cleansing step of our approach, we choose the most relevant tags based on their frequency and their meaningfulness, as described in section 3.1. This resulted in a data set of roughly 18,329 images. The associated vocabulary has well over 6,798 terms chosen by 1,904 users.

We used also data from the social bookmarking system Del.icio.us, collected in November 2006. In total, data comprise 663, 950 users, 2, 398, 483 tags, and 18, 775, 470 resources. After the cleansing the resulted dataset has roughly 23, 643 users, 48, 135 resources, and 8, 568 tags.

B. Preparation of tags

In this step, we generate a list of tags ordered by their generality degree. The generality measure implemented is the proposed one (FDU). Table 5 and 6 depict excerpts of this list for Flickr and delicious dataset respectively.

[3] http://www.flickr.com.

Table 5. An excerpt of generality list of Flickr dataset measured by FDU

Tag	Generality degree
Food	4561
Music	1114
Sport	856
Restaurant	373
Japan	366
Transport	276
Friend	265

Table 6. An excerpt of generality list of Delicious dataset measured by FDU

Tag	Generality degree
Web	8680
Design	4720
Html	2950
Blog	2373
Search	2280
News	2200
Art	1290

Table 7. An excerpt of the NCDU Matrix

	Food	Girl	Airline	Music	Sport
Actress	0,00	0,67	0,00	0,33	0,00
Adventure	0,20	0,00	0,00	0,00	0,40
Advertise	0,58	0,00	0,00	0,08	0,08
Airplane	0,61	0,00	0,50	0,05	0,05
Banjo	0,00	0,00	0,00	0,71	0,00
Baseball	0,10	0,02	0,00	0,00	0,86
Guitar	0,00	0,02	0,00	0,83	0,01

Table 8. Most similar 4 terms to some tags using NCDU

	(1)	(2)	(3)	(4)
Architecture	Rio	Skyscraper	Historic	Building
Canada	Quebec	Edmonton	Ontario	Toronto
Art	Actress	Rio	Craft	Fine
Fun	Food	Lifestyle	Teenager	Zoo
Cat	Kitten	Food	Dolphin	Amazon
Band	Music	Ballroom	Gig	Guitarist
Famous	Actress	Food	Interestingness	Celebrity
Car	Nissan	Automotive	Ford	Mustang

Table 9. Most similar 4 terms to some tags using Co-occurrence

	(1)	(2)	(3)	(4)
Architecture	Animal	Flower	Cat	Tree
Canada	Food	Animal	Flower	Cat
Art	Animal	Music	Flower	Tree
Fun	Music	Animal	Friend	Cat
Cat	Animal	Flower	Tree	Dog
Band	Music	Live	Rock	Nightclub
Famous	Food	Yahoo	Interestingness	Brazil
Car	Day	Race	Bahrain	Animal

Table 10. Most similar 4 terms to some tags using Co-occurrence

	(1)	(2)	(3)	(4)
Architecture	Red	Fish	Travel	January
Canada	Photo	Fish	Doughnut	Drink
Art	Tokyo	June	Topv	Reflection
Fun	People	July	Team	Beer
Cat	House	France	Grill	Cookie
Band	Travel	China	Macro	Birthday
Famous	Scott	South	Brighton	Dance
Car	April	Travel	Red	Birthday

In this step, we generate also a similarity matrix of our folksonomy based on the NCDU measure. This similarity measure gives more accurate results than CDU, co-occurrence, and cosine. As an example, we give in table 7 an excerpt of the NCDU matrix and in Table 8-10 we give the most similar 4 terms to some tags.

C. Preparation of tags

As explained above, context identification and disambiguation are performed in one-step by applying a new clustering algorithm. Fig. 3 and 4 depict examples of ambiguous tags “player”, and “adventure” from Flickr. Fig.3 shows an example of the ambiguous tag “design” as they are detected by the algorithm.

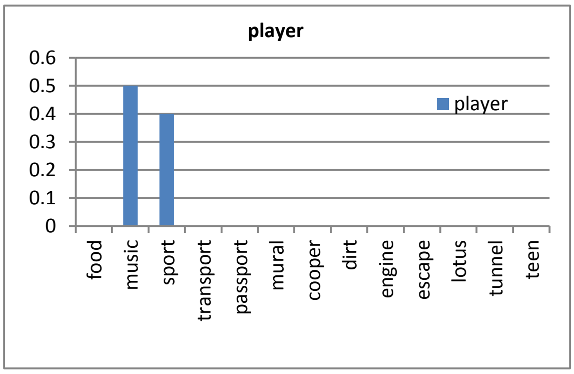

Fig. 3. The ambiguous tag ”player” as detected by the algorithm

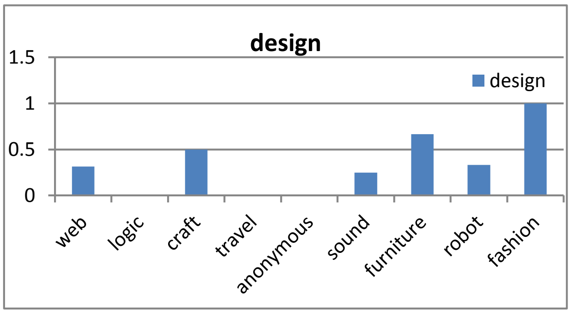

Fig. 4. The ambiguous tag ”design” as detected by the algorithm

Table 11 shows other examples of ambiguous tags detected by this algorithm.

Table 11. An expert of discovered ambiguous tags with their context

Tags	Cluster centers
Track	Sport / transport

Entertainment	Music/sport
Swing	Music/sport
Speed	Sport/ transport / engine
Music	Mp/ loop/ sound
Spider	Web/ robot
Video	Multiplication/ sound

D.Semantic identification

In this step, we have generated hierarchies of tags using the algorithm described in Section 3.4. Fig. 5. and 6., illustrate excerpts of these hierarchies for Flickr and Delicious tags respectively.

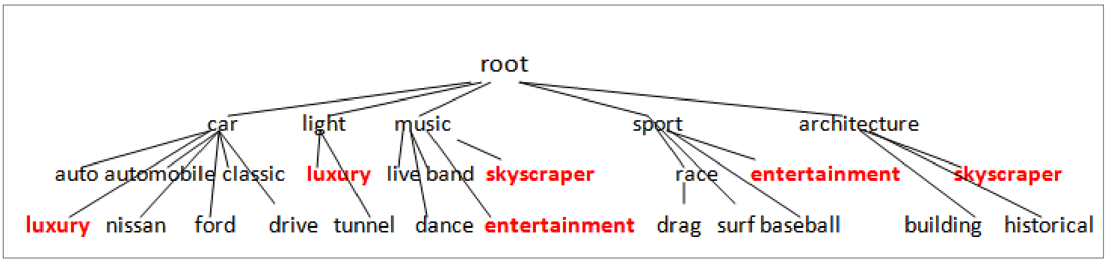

Fig. 5. A small excerpt from the tag hierarchy created for Flickr tags (red tags are ambiguous ones)

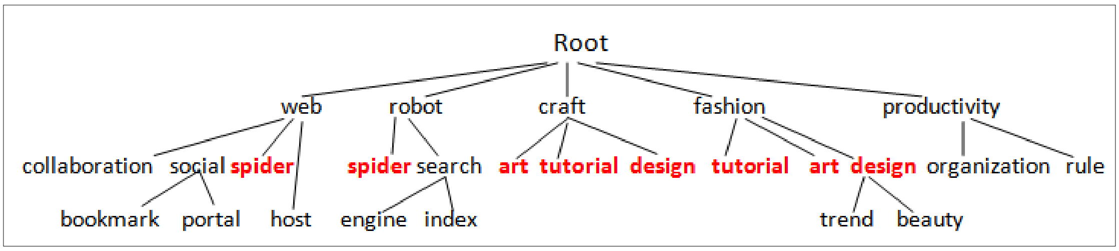

Fig. 6. A small excerpt from the tag hierarchy created for Delicious tags (red tags are ambiguous ones)

E. Evaluation

In order to evaluate the quality of the learned hierarchy, we compare it with manually built categorization schemes from WordNet and Wikipedia. Despite that, it is not obvious to find a valid similarity score for two hierarchical structures. For this purpose, we use the measures proposed by [8], namely F-measure and taxonomic overlap measures. The principal is to find a concept present in the two hierarchies and to extract excerpt from both ontologies containing this concept, then the similarity of the two hierarchies depends of the similarity of the two excerpts.

Based on these measures and both ontologies we run several experiments to assess the quality of our approach. For this purpose, we have generated taxonomies based on co-occurrence, cosine, and NCDU measures. The results of the comparison are depicted in Fig. 7 based on taxonomic overlap, and in Fig. 8 based on F-measure.

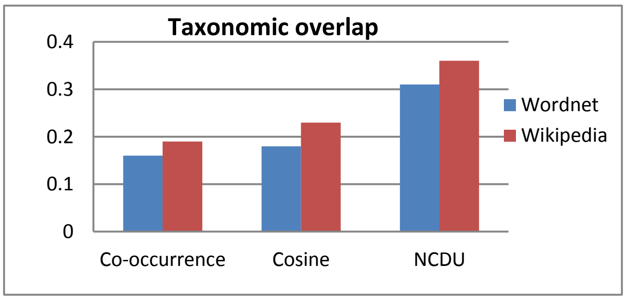

Fig. 7. Taxonomic overlap based comparison between the learned hierarchical structures and the reference ontologies from WordNet and Wikipedia

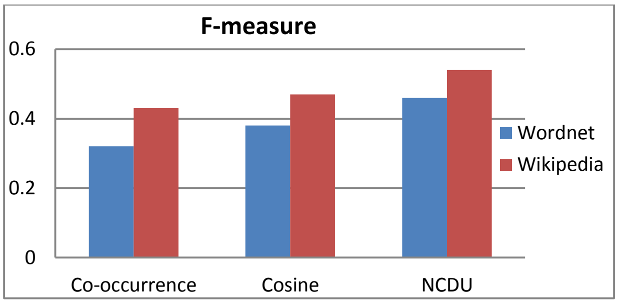

Fig. 8. F-measure based comparison between the learned hierarchical structures and the reference ontologies from WordNet and Wikipedia.

We have also tested the quality of the proposed similarity measure (NCDU). For this reason, we have used Kendall rank correlation coefficient referred to as Kendall's τ coefficient. It is a statistic used to measure the association between two measured datasets.In our experiments, we calculate τ correlations between the similarity values based on co-occurrence, cosine and NCDU measures on one hand and the reference similarity values provided by the WordNet grounding measure on the other hand. The formula of τ is as follows:

$$\tau = \frac{\begin{array}{c}(\text{number of concordant pairs})\\ -(\text{number of discordant pairs})\end{array}}{\frac{1}{2}\,\mathrm{P}\,(\mathrm{P}-1)}$$

Where P is the total number of pairs, and concordant pairs are defined as follows:

Given two pairs p1= (t_1; t_2) and p2= (t_3 ; t_4). p1 and p2 are concordant if $t_1 > t_3$ and $t_2 > t_4$, or if $t_1 < t_3$ and $t_2 < t_4$. Values of τ ranges from -1 to 1, where -1 means that all the pairs are discordant, and 1 means that all are concordant.

Fig. 9 plots Kendall's correlation between generated similarities, and the WordNet reference.

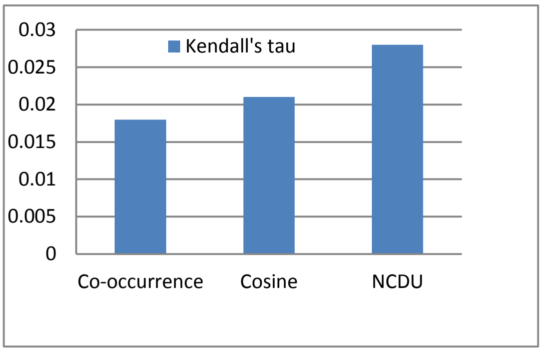

Fig. 9. Experimental results of comparing generated similarities with WordNet reference

We have used the same formula to compare the proposed generality measure FDU with the frequency based generality measure. However, here we have another definition of concordant pairs and discordant ones. Given a pair of tags p= (t_1; t_2), p is a concordant pair if we have $t_1 > t_2$ in a generality measure as well as in WordNet.

Fig. 10 illustrates comparison between these generality measures.

The objective of evaluating a fuzzy clustering algorithm is to seek clustering schemes where most of the vectors of the dataset exhibit high degree of membership in one cluster. For this purpose, we have used the partition coefficient (PC) [11] to compare our algorithm to the FCM.

$$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} {u_{ij}}^2$$

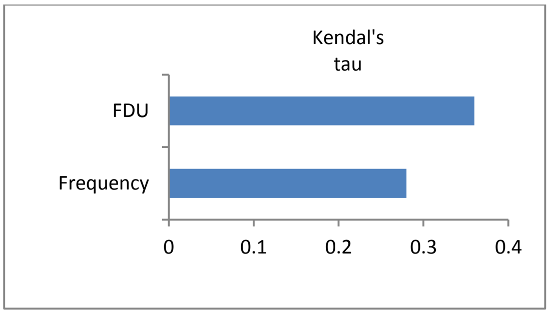

Fig. 10. Experimental results of comparing generality degrees with WordNet reference.

Where *n*, and k *are* tags count and clusters count respectively.

We perform a set of experimental results to compare the performance of our new algorithm and the FCM. The experiments are done on a personal computer with Intel core i3 2,4 GHz at 64 bit, 4 GB of memory and 464 GB hard disk. It is obvious that our new algorithm outperforms in a remarkable way the FCM algorithm in term of quality of clusters and in term of performance as shown in Fig. 11 and Table 12.

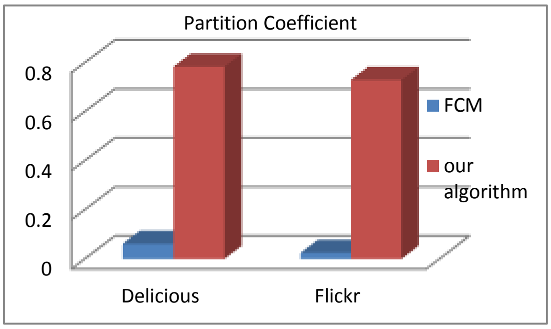

Fig. 11. Experimental results of comparing the quality of our new algorithm and the FCM one based on the partition coefficient PC.

Table 12. Experimental results of comparing the runtime performance of our new algorithm and the FCM.

	Delicious	Flickr
FCM's runtime in seconds	622,349	593,288
Our algorithm's runtime in seconds	1,214	1,035

V. Conclusion

Folksonomies have their own shared vocabularies and relations, which can be extracted as an ontological structure. In recent years, there has been a growing number of research works trying to associate semantic information to tags in folksonomies. The main objective of these works is to identify the shared conceptualizations hidden in folksonomies.

Although several approaches have been proposed to bring structure to folksonomies, they do not come without limitations. These include the inability to decide the rules generated by association rule mining as to which term is more general or narrow, and tags that cannot be found in the upper ontologies. Moreover, tags disambiguation task and calculus of similarity between tags still suffer from several limitations.

In this paper, we have proposed an integrated approach to extract ontological structures from collaborative tagging systems. We proposed a simple representation of the folksonomy in the vector space model to reduce calculations time. Our approach tries also to overcome the limitations of the other approaches by introducing all levels of a folksonomy in calculating similarity between tags. We propose also anew algorithm to define context of tags and to detect ambiguous ones. Moreover, we ameliorate the approach of Heymannet al [7] by employing the clusters generated in the context identification and disambiguation task, in the hierarchy generation task. The study shows the approach has significant potentials for ontology extraction from folksonomies.

For future work, we will attempt to ameliorate the similarity measure used in this work by combining it with semantic similarity measures like those presented in [33].

Moreover, we plan to improve the approach by discovering non-taxonomic relations by detecting tags representing verbs and their related tags. Finally, we plan to extrinsically assess the quality of our results by integrating them in the context of various tasks such as tag disambiguation, result visualization, and ontology evolution.

REFERENCES

[1] Zimmer, M., Preface: ‘Critical Perspectives on Web 2.0." First Monday, 13(3), 2008.

[2] Vander, T: Folksonomy Coinage and Definition, http://www.vanderwal.net/folksonomy.html.

[3] Mika. P., Ontologies are us: A unified model of social networks and semantics. In International Semantic Web Conference, LNCS, pages 522–536. Springer, 2005.

[4] Hamasaki, M., Matsuo, Y., Nisimura, T. & Takeda, H. Ontology Extraction using Social Network. In International Workshop on Semantic Web for Collaborative Knowledge Acquisition, India, 2007.

[5] Begelman G., Keller P., & Smadja F. Automated tag clustering: Improving search and exploration in the tag space. In 15th International World Wide Web Conference, Edinburgh, Scotland, 2006.

[6] Kennedy L., Naaman M., Ahern S., Nair R., & Rattenbury T. 2007 How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections. In Proceedings of ACM Multimedia, Augsburg, Germany. 2007.

[7] Heymann. P.,and Garcia-Molina. H. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department, April 2006.

[8] Benz, D., Hotho, A., Stumme, G.: Semantics Made by You and Me: Self-emerging Ontologies can Capture the Diversity of Shared Knowledge. In: Proceedings of the 2nd Web Science Conference (WebSci10), USA. 2010.

[9] Marouf, Z., Benslimane, S. M. fuzzy clustering-based approach to derive hierarchical structures from folksonomies. International conference on computer systems and applications, AICCSA 2013, Maroc.

[10] Maedche, A., &Staab, S. Ontology Learning for the Semantic Web. IEEE Intelligent Systems 16(2), 72-79. 2001.

[11] Bezdek, J.C. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York. 1981.

[12] Schmitz, C., Hotho,A., Jaschke, R., and Stumme .G.: Mining association rules in folksonomies. In the Proc. of the 10th IFCS Conf., Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270, Berlin, Heidelberg, 2006.

[13] Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G., Discovering shared conceptualizations in folksonomies. In Journal of Web Semantics 6(1), 38-53. 2008.

[14] Lehmann. F., Wille. F., R., A triadic approach to formal concept analysis, in: G. Ellis, R. Levinson, W. Rich, J. F. Sowa (eds.), Conceptual structures: applications, implementation and theory, vol. 954 of Lecture Notes in Artificial Intelligence, Springer Verlag, 1995.

[15] Trabelsi. C., Jelassi. N. and Ben Yahia. S. Scalable Mining of Frequent Tri-concepts from Folksonomies. The 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining,pp 231-242, 2012.

[16] Angeletou, S., Sabou, M. & Motta, E. Semantically Enriching Folksonomies with FLOR. In 1stInternational Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008), Tenerife, Spain. 2008.

[17] Cantador, I., Szomszor, M., Alani, H., Fernandez, M. & Castells, P., Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations. In 1st International Workshop on Collective Semantics, (CISWeb 2008), Tenerife, Spain, 2008

[18] Garcia-Silva, A., Szomszor, M., Alani, H., &Corcho, O., Preliminary Results in Tag Disambiguation using DBpedia. In 1st International Workshop in Collective Knowledge Capturing and Representation (CKCaR09), California, USA, 2009.

[19] Auer S., Bizer C., Kobilarov G., Lehmann. J., Cyganiak R., and Ives. Z.,DBpedia: A Nucleus for a Web of Open Data. 6th International Semantic Web Conference, 2007.

[20] Djuana, E., Xu, Y., Li, Y., Learning Personalized Tag Ontology from User Tagging Information. Conferences in Research and Practice in Information Technology (CRPIT), Australia, 2012.

[21] Giannakidou, E., Koutsonikola, V., Vakali, A., &Kompatsiaris, Y. Co-Clustering Tags and Social Data Sources. In Proc. 9th International Conference on Web-Age Information Management, 2008.

[22] Specia, L., Motta, E.: Integrating Folksonomies with the Semantic Web. In: 4th European Semantic Web Conference, pp. 624-639. 2007

[23] Lin, H., Davis, J. and Zhou, Y., An integrated approach to extracting ontological structures from folksonomies. In Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications, page 668. Springer, 2009.

[24] Schmitz, P.: Inducing ontology from Flickr tags. Collaborative Web Tagging Workshop, 15th WWW Conference, Edinburgh, 2006.

[25] Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., Stumme, G.: Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In: 18t International Conference on World Wide Web, pp. 641-650. 2009.

[26] Angeletou, S., Sabou, M., Motta, E.: Improving Folksonomies Using Formal Knowledge: A Case Study on Search. In: 4th Asian Semantic Web Conference, pp. 276-290. 2009.

[27] Weinberger, K. Q., Slaney, M., Van Zwol, R.: Resolving Tag Ambiguity. ACM Multimedia, page 111-120. ACM, (2008).

[28] Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R. 2008. Personalized Recommendation in Social Tagging Systems using Hierarchical Clustering. ACM Conf. on Recommender Systems, pp 259-266. 2008.

[29] Au Yeung, C. M., Gibbins, N., Shadbolt, N.: Contextualising Tags in Collaborative Tagging Systems. In: 20th Conference on Hypertext and Hypermedia, pp. 251-260. 2009.

[30] Pantel, P., and Lin, D,. Document clustering with committees. In Proc. of SIGIR’02, Tampere, Finland, 2002.

[31] Strohmaier, M. Helic, D. Benz, D. Körner, C. and Kern. R. Evaluation of folksonomy induction algorithms. ACM Trans. Intell. Syst. Technol., 2012

[32] Hoser, B., Hotho, A., Jaschke, R., Schmitz, C., Stumme G., Semantic network analysis of ontologies. In European Semantic Web Conference, Budva, Montenegro, June 2006.

[33] Kavitha, A., Rajkumar, N., and Victor, S.P., An Integrated Approach for Measuring Semantic Similarity Between

Words and Sentences Using Web Search Engine. The International Journal of Information Technology & Computer Science (IJITCS), 9(3), 68-78.2013.

Authors' Profiles

Zahia Marouf is an Assistant Professor at Faculty of Economics, business studies and management of Mascara University, Algeria. She received her magister degree in computer science from Mascara University in 2010. She also received an engineer degree in computer science in 2006 from the Computer Science Department of Mascara University. Here research interests include collaborative tagging systems, semantic web, ontology engineering, information and knowledge management.

Sidi Mohamed Benslimane is an Associate Professor at the Computer Science Department of Sidi Bel Abbes University, Algeria. He received his PhD degree in computer science from Sidi Bel Abbes University in 2007. He also received a M.S. and a technical engineer degree in computer science in 2001 and 1994 respectively from the Computer Science Department of Sidi Bel Abbes University, Algeria. He is currently Head of Research Team 'Service Oriented Computing' at the Evolutionary Engineering and Distributed Information Systems Laboratory, EEDIS. His research interests include, semantic web, service oriented computing, ontology engineering, information and knowledge management, distributed and heterogeneous information systems and context-aware computing.

How to cite this paper: Zahia Marouf, Sidi Mohamed Benslimane,"An Integrated Approach to Drive Ontological Structure from Folksonomie", International Journal of Information Technology and Computer Science(IJITCS), vol.6, no.12, pp.35-45, 2014. DOI: 10.5815/ijitcs.2014.12.05