

When Hashtags Meet Recommendation in e-learning Systems

Mérième GHENNAME

¹Laboratoire Hubert Curien, Université Jean Monnet 42000, Saint-Etienne FRANCE.

²Université Mohammed V de Rabat 10000, Rabat MOROCCO.

ghenname.merieme@gmail.com

Julien SUBERCAZE

¹Laboratoire Hubert Curien, Université Jean Monnet 42000, Saint-Etienne FRANCE.

firstname.lastname@univ-st-etienne.fr

Frédérique LAFOREST

¹Laboratoire Hubert Curien, Université Jean Monnet 42000, Saint-Etienne FRANCE.

firstname.lastname@univ-st-etienne.fr

Mounia ABIK

²Université Mohammed V de Rabat 10000, Rabat MOROCCO.

name@ensias.ma

Christophe GRAVIER

¹Laboratoire Hubert Curien, Université Jean Monnet 42000, Saint-Etienne FRANCE.

firstname.lastname@univ-st-etienne.fr

Rachida AJHOUNE

²Université Mohammed V de Rabat 10000, Rabat MOROCCO.

name@ensias.ma

Abstract: Knowledge, education and learning are major concerns in today's society. The technologies for human learning aim to promote, stimulate, support and validate the learning process. Our approach explores the opportunities raised by mixing the Social Web and the Semantic Web technologies for e-learning. More precisely, we work on enriching learner's profiles from their activities on the social Web. We propose a methodology for exploiting hashtags contained in users' writings for the automatic enrichment of learner's profiles. This paper aims at giving an insight on the processing required on hashtags before being source of knowledge on the user interests. For this purpose we introduce our approach for the automatic structuration of hashtags definitions into synonym rings. We present the output as a so-called folksonary, i.e. a single integrated dictionary built from everybody's definitions. Semantized hashtags are thus used to feed the learner's profile and particularly the focus field.

Keywords: Hash tags, Social Network, Semantic Web, E-learning, Learning Profile, Personalized Recommendation.

1. PROBLEM DEFINITION

The development of the Internet has greatly facilitated the learning. We have seen the emergence of platforms that help users to learn at their own place and with respect to their availability. Those systems participate to increase the involvement of learners and decrease their feelings of isolation. But Despite all their advantages the majority of Learning Management System (LMS) platforms focus on publishing pedagogical materials to the learner, and are almost little interested to the personalization of recommended content. The way we learn is motivated and guided by the emergence of new Web technologies, learning becomes, contextual, personal, and collaborative. Actually LMS attempt to make the learning process more effective and relevant. However effective and relevant means to consider several parameters, such as profile's learners, their needs, and their focus. In other words the system must be able to determine which resources are required to the learner profile. Consequently we instantiate our problem of recommendation based profile, on discovering learner's profiles problem. Thus we see the social Web as a key way to gain more information about the learners' profiles and the semantic Web as an effective means to structure and make usable the information derived.

The social Web connects and captivates the attention of millions users. As a consequence users' writings and data on social networks are growing exponentially over time, and they become hardly exploitable. So hashtags have become a lightweight solution to classify and search information on the Web 2.0 and 3.0. Unfortunately, a hashtag is at best a composed word, and at worst a neologism. It is

not a piece of information by itself. The primary information is the association (the tagging relation) that exists between a hashtag and a resource. It is however important to gain more knowledge on hashtags. In the literature, this problem is usually addressed using the context involving a hashtag [3], [2]. For instance in a textual resource, this context is the terms window surrounding the hashtag. While this provides an information source to enrich hashtags with related bag of- words, it suffers from several drawbacks. Firstly, the tag context is noisy, given that a tag may have been associated to hundreds of thousands or even millions resources. Moreover, the tremendous amount of data makes it impossible to retrieve all resources associated to a given hashtag. These two issues make the context of use of a hashtag incomplete and noisy. Enriching tags has not only become a knowledge discovery issue, but it is also a problem for the end-user. So, end-users feel the need to define the hashtags they use, for reusability and explanation purpose. For this, several web services are available so that any user can publish her own definition of a hashtag. One can cite *Tagdef.com* or *Hashtags.org*.

Our intuition is that such crowd-sourced services may turn out to be interesting sources of information to gain knowledge on hashtags. However their lack of structure compared to usual external databases used in IR or NLP (mainly Wordnet or DBPedia¹) restricts the scope of possibilities offered by the service. We attempt to introduce a first structuration of crowdsourced hashtags dictionaries using state of the art NLP and clustering techniques. If one refers to WordNet² he could expect to introduce the same kind of relationships such as super-subordinate (hypernymy, meronymy, synonymy. . .).

2. RELATED WORK

The major concern of today's e-learning systems is to upgrade their skills and capabilities. As well the development of Web technology plays an ever more central role in the improvement of e-learning environment and affect significantly the methods, and disciplines of learning. Because on one hand the collaborative aspect of the social Web contributes effectively to enhance and extend the functionality of platforms as e-learning. And on the other hand the semantic aspect of Web 3.0 enable a better structuring of information and the opportunity to combine various heterogeneous information from different sources. In the context of our work we focus on the social contributions mainly hashtags. In this perspective, we have conduct a review of previous pioneers' works for enhancing learning when both the social and the Semantic Web are involved in the approach. Recently, researchers and developers in learning technologies have started to combine Social Web and Semantic Web techniques. Both paradigms aims at giving a well-defined meaning for information, and opportunities such as learning individualization, free knowledge access, the opening of training to new and wider audience despite the distance of any kind (geographic, cultural, social, economic and transactional).

We can quote here [5], [6] where authors suggest how developments in semantic technologies can be used to generate personalized learning environments that will motivate learners. They describe on one hand a Semantic Web-based e-learning architecture and the importance of using metadata in the e-learning field. On the other hand, they list the challenges of incorporating the semantic Web in a learning process and which may be essentially summarized in : achieving interoperability between different educational systems, automate the process of knowledge creation and the structuring and the unification of educational data. As well as to open, share and reuse the content of education systems and knowledge components. Moreover, collaborative tagging has grown in recent years, with sites that allow users to tag bookmarks, photographs and other content. Consequently, following the tagging practice many researchers realized the need to concoct the semantic Web to social Web, and many of them have focused on approaches for the semantic disambiguation of hashtags. Some research as [7] has introduced MOAT a lightweigh a collaborative framework which goal is to let Users Bridge the gap between free-tagging and semantically annotated content in a simple way. Their approach relies on Linked Data principles to enable people to relate content with any URI from existing resources, in order to let their content enter the Semantic Web and at the same time solve the limits of free-tagging. We also quote [8] where authors exhibit a methodology based on online lexical resources, ontologies and Semantic Web resources for to enrich hashtags with semantics with the aim to integrate folksonomies and the semantic Web. And [9] extending the approach of [8] they also offers an interesting and comprehensive approach for semi-automatic generation of ontologies out of

¹ <http://dbpedia.org>

² <http://wordnet.princeton.edu/>

folksonomies. But they also involve the community as a mechanism to validate all the information extracted from the resources. Hence they propose to involve human intelligence as a community approval of the resulting conceptualization, to confirm the semantics obtained from existing ontologies and resources.

More contributions have worked on leveraging the above technologies in the e-learning domain. For example, DERI Galway has developed a framework for extracting useful knowledge published online in an informal way (e.g. wikis, blog posts, forum posts), structuring the acquired knowledge and putting it into use within LMSs [10]. For their part the authors of [11] and propose a collaborative semantic-rich learning environment in which folksonomies created from students' collaborative tags contribute to ontology maintenance, and teacher-directed feedback. We also found [12] proposing a method for the formulation and interpretation of learning management platforms as social networks. Thus they develop an ontology to integrate the information from different Learning Management Systems, and from which a personalized social network is extracted. This vision allows to make further studies about learners, teachers and learning resources to obtain a better understanding of their social structure, and therefore to make or improve decisions about the learning process. Otherwise the contribution of [13] relies on the Social Semantic Web paradigm to prove learning systems and tools and provide students with context-aware learning services. It's developed using active learning techniques, project-based learning, and collaborative learning. The underlying philosophy of DEPTH is based on the fact that the major concern of today's software engineering education is to provide students with necessary skills to integrate theory and practice; to have them recognize the importance of modelling and appreciate the value of a good design; and to provide them with the ability to acquire specific domain knowledge, in order to support software development in specific domains.

As well through review of previous work, we identify how the social and semantic Web can improve the semantic richness of hashtags, and could also be applied for analysis of learner's hashtags in order to increase their profile shared content: annotated resources and/or tags used for annotation. Our vision intended to exploit communicative dimensions of social Web, reactivity and the engagement of users on social structures, to overcome the ambiguity of hashtags. Furthermore the semantic enrichment of hashtags allow us to build a rich hashtag dictionary, containing maximum hashtag with their definitions on one hand. On the other hand to a rich database on which we can do treatments algorithmic and deduce the users interests. The result will be a structured vocabulary ensuring better visibility of users and enable to increase their learning profile. We fully discuss our approach in the following sections.

3. APPROACH FOR FOLKSIONARY

In this section we present an approach that provides a clustering of user-generated definitions into different senses, over any dataset of words along with their user-generated definitions in natural English.

3.1 Formalization

let w be a set of words. for each $w \in W$ we define $d(w)$ the set of definitions for w and $s(w)$ the set of possible senses for w . we denote $dw_{w,i}$ the i -th definition of the word w , where $i \in [1; |d(w)|]$.

we use the function *employed as* denoted ε that relates each definition of a word to a sense of the same word as follows:

$$\varepsilon : D(w) \rightarrow S(w) \tag{1}$$

$$\forall d \in D(w), \exists s \in S(w), \varepsilon(d) = s$$

Definition 1: Function ε is a surjective function (c.f. Equation 1), therefore every sense of every word in the dictionary consists of at least one user-generated definition.

Definition 2: $S(w)$ is a partition of $D(w)$, such as every user-generated definition of a word w belongs to exactly one sense $s \in S(w)$, which means :

$$\begin{cases} \cup s_{w,i} \in S(w) = D(w) \\ \forall s_1, s_2 \in S(w), s_1 \neq s_2 \rightarrow s_1 \cap s_2 = \emptyset \end{cases} \quad (2)$$

We formalize the similarity matrix $Dist(w)$, with a normalized matrix which expresses the distances, taken pairwise, of a set of definitions for a given hashtag:

$$Dist(w) = (dist(d_{w,i}, d_{w,j}))_{1 \leq i, j \leq |D(w)|} \quad (3)$$

3.2 Process for Building A Folksionary

To build a folksionary, we perform a four-steps process. First, we crawl hashtags definitions from online services. Secondly, for each hashtag, we perform a pairwise comparison of its definitions by computing a distance between pairs of definitions. At third step, we apply a clustering algorithm for each hashtag in order to group its definitions into similar meaning clusters. Lastly, we export these results under the form of a human-readable document with a look very close to a standard dictionary. Figure 1 illustrates this approach. The following sections, detail these four steps.

3.2.1 Crawl hashtags definitions: This step populates W as well as $D(w), \forall w \in W$.

Different sources of data on the Web contain users written definitions of hashtags in natural language. For instance, *Tagdef.com* or *Hashtags.org* are well-known online hashtags dictionaries. In first step, we crawl hashtags and their definitions from such sources. The scrapping process extracts definitions from each given page and, using a language classifier keeps only english definitions.

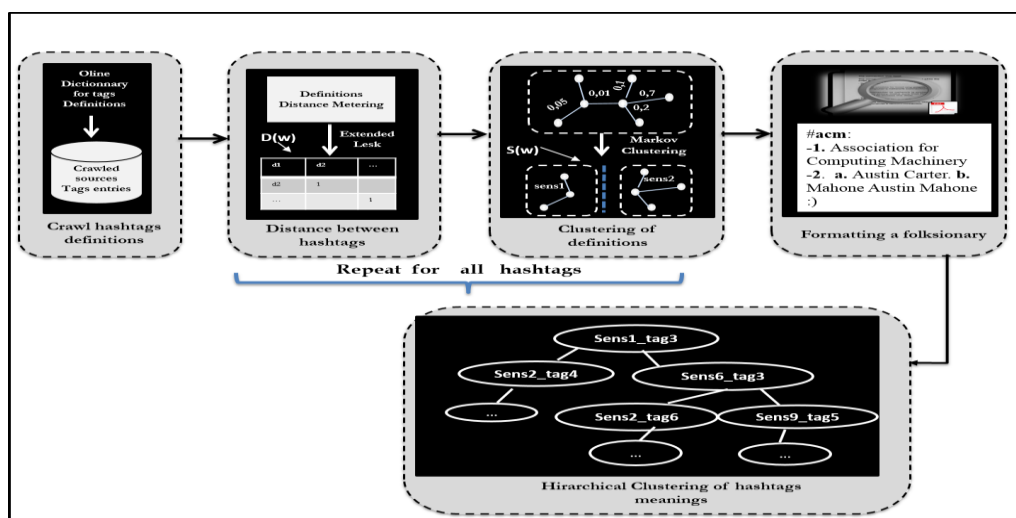


Fig1. Approach for building a folksionary and hierarchical clustering

$\forall w \in W$

3.2.2 Distance between hash tags: The objective of this step is to populate $Dist(w)$,

User-generated definitions for a given hashtag can be redundant, i.e. some definitions can describe the same meaning. Our goal in this step is to measure the semantic-relatedness between definitions for the clustering phase (section III-B3). In the literature, the traditional approach to compare two sentences relies on the co-occurrence frequency of terms employed in the different natural language sentences [15], [16]. These approaches are limited to the strict co-occurrence of the same terms in the definitions. But crowd-sourced hashtags definitions are populated by different users, using heterogeneous terms, neologisms and abbreviations. We need an external knowledge base, to take into account proximity between terms in the metric between hashtags definitions. This issue is referred as semantic-relatedness for the word sense desambiguation problem [17]. Among techniques involving an external knowledge base, the Extended Lesk algorithm has proven to be one of the most efficient [18]. Extended Lesk is an adaptation of the Lesk [19] using Wordnet3 as an external knowledge base.

³ <http://wordnet.princeton.edu>

Using the context of use (a term window) of a given target word, it selects the most plausible sense for this word from all the possible senses in Wordnet. This algorithm is limited to the semantic-relatedness between two words. [20], propose a new approach for the semantic-relatedness between two sentences using Extended Lesk. We use Extended Lesk on each set $D(w)$ to provide the semantic-relatedness between definitions of the hashtag w under the form of a matrix. Each matrix represents the adjacency matrix of a weighted graph where edges are the definitions of a hashtag, and the vertices are weighted by the distance between the two definitions.

3.2.3 Clustering of definitions: The objective of this step is to populate $S(w)$, $\forall w \in W$.

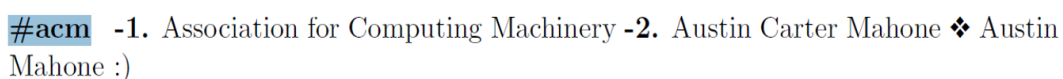
In the previous step we generate a graph providing $\text{Dist}(w)$, the distances between definitions of a hashtag. This graph is used to cluster hashtags based on their meanings. In our approach we have no a priori information regarding the number of clusters. A comparative analysis has shown that the Markov Clustering algorithm (MCL) [21] is remarkably more robust than other clustering techniques [22]. It produces good clustering results mainly because the algorithm scales well with increasing graph size, it is robust against noise in graph data even if it cannot find overlapping clusters. Also we are not constrained to specify a number of clusters beforehand. MCL interprets the matrix entries or graph edge weights as similarities. It simulates a random walk in the graph by changing iteratively the transition probabilities in adjacency matrix with normalized value in $[0;1]$. In MCL two processes alternate: Expansion and Inflation. The expansion operator connects different regions of the graph, and the inflation operator is responsible for strengthens and weakens. Eventually, iterating expansion and inflation results in the separation of the graph into different segments. The collection of resulting segments is simply interpreted as a clustering. Several parameters are available for tuning the mcl computing process. The most popular are inflation parameter setting for obtaining clusterings at different levels of granularity, the measure of idempotence and pruning, the maximum value considered zero for pruning operations and values for cycles. During this step, for each hashtag, we group its definitions into units of meaning $S(w)$. We are then able to perceive to what extent each hashtag is polysemic with the cardinality of $S(w)$.

3.2.4 Formatting a folksionary

One of the objectives of a folksionary is to provide a new kind of dictionary to human users. Therefore we output the in-memory model of the Folksionary in a format close to a traditional dictionary. This output is a PDF file that organizes hashtags entries in an alphabetic order. Each hashtag is presented with all its meanings, and we list in each meaning all the definitions that were clustered. For instance, consider the following definitions that were crawled for the hashtag #acm at step 1 (c.f. III-B1):

- Austin Carter Mahone”
- “Association for Computing Machinery”
- “Austin Mahone :)”

We present it using a standard dictionary formatting (see figure 2):



```
#acm -1. Association for Computing Machinery -2. Austin Carter Mahone ❖ Austin Mahone :)
```

Fig2.A standard dictionary formatting

As shown in the previous example, two meanings were detected for the hashtag acm, one for Association for Computing Machinery (with one definition) and the other for the person named Austin Carter. This second meaning comes from two different definitions, which were grouped in the same cluster. The different symbols are intended as the following :

- The different sense $s \in S(w)$ are separated by numbers. -1. denotes the first meaning, -2. denotes the second S meaning, and so on.
- Definitions of the same sense $\forall d_i \in s$ with $s \in (w)$ are separated by \diamond

3.3 Hierarchical Clustering

In this section we carried out a hierarchical clustering on our folksionary, to explore relations between hashtags. Since this clustering outputs an informative structure about relations between objects. We opted for an Ascending Hierarchical Clustering algorithm (CHA) [34]. It's works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. We used the same distance measurement in 3.3 as it was already tested and gives good results on our data. Whereas to merge clusters we rely on average linkage (average-average distance). We perform an Ascending Hierarchical Clustering between all hashtags meanings on the folksionary. And basis of dendrogram graph we model the output of clustering. The CHA allowed us to have more visibility on the relations between hashtags on the folksionary. We concentrated mainly on synonymy in this work, other relations are to discover in a future works. Hashtags are ambiguous, so to discover relations between two hashtags, we must discover relations between their meanings. In others words we calculate similarity between the meaning of the two hashtags and the we apply the CHA. This will be repeated for all hashtags of the folksionary. Consider the example hereafter, there is a relation between pine and cone. More precisely the first meaning of **#pine** is equivalent to the third meaning of **#cone**.

- **#pine** -1. kinds of evergreen tree with needle-shaped leave -2. waste away through sorrow or illness.
- **#cone** -1. solid body which narrows to a point -2.something of this shape whether solid or hollow -3. fruit of certain evergreen trees.

Fig. 3 represents an output of the hierarchical clustering algorithm applied to our folksionary. The height of a cluster in the dendrogram is equal to the similarity between two clusters before merging, while the leaves give the meanings of hashtags. The right scale is the hierarchy index. The name of each meaning is composed of the hashtag's name and its count in the hashtag's meanings. In the example we have three hashtags 100wordchallenge, 100factsaboutme and 10212011.

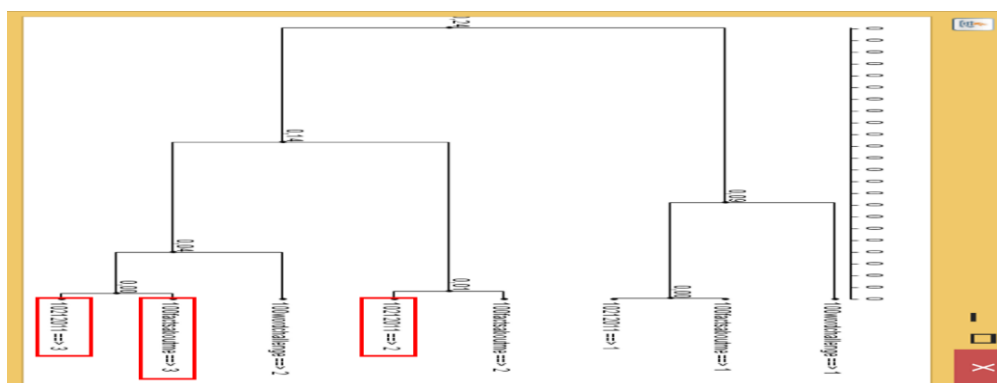


Fig3.Hierarchical clustering on the folksionary

4. PROTOTYPE AND EVALUATION

This session is dedicated to our prototype and the characterization of results obtained on the folksionary. We also provide a qualitative analysis measuring the distance between the generated folksionary and a ground truth established manually

4.1 Prototype implementation

To demonstrate our approach, we have constructed a dataset by crawling web sources. For this purpose, we have created dedicated Web scrappers using the pjsrcrape javascript library⁴. It performs a browser-like rendering, therefore we did not miss any AJAX-generated content. We used Apache Tika [23] for language filtering in order to select only English definitions. Then, we compute the distance using an in-house developed Java version of Extended Lesk. And we finally perform Markov Clustering using JavaML [24]. In this section we detail the characteristics of our folksionary and provide an evaluation.

⁴ <http://nrabinowitz.github.com/pjsrcrape/>

4.2 Folksionary Characterization

We have built a folksionary by applying our approach on the aforementioned dataset. The folksionary PDF file containing all tags is available online at: <http://goo.gl/1b2Jp8> This folksionary contains 22,738 hashtags, and a total of 28,191 definitions. Our approach identified 25,106 meanings in 28,191 definitions. Each hashtag has an average of $\sim 1; 1$ meaning (SD : $\sim 0,45$). In this folksionary, 1,731 hashtags out of 22,738 have several meanings.

Let us focus on the 1,731 tags that have been detected polysemic. Polysemic hashtags have on average $\sim 2:37$ meanings with a standard deviation of $\sim 0; 94$. Figure 3 presents the number of tags grouped by number of meanings. For instance: 261 hashtags have three distinct meanings detected by our approach. 98,8% hashtags are polysemic with at maximum five different meanings. The last 1,2% tags with a degree of polysemy superior or equal to 6, represent a tiny portion of our folksionary and are considered as exceptions in this work. Those tags are hugely popular tags, such as #justinbieber where people express different, sometimes ironical definitions.

4.3 Evaluation

In order to complete the quantitative analysis of our folksionary, a qualitative analysis is needed. It consists in measuring the distance between the generated folksionary and the Ground Truth. The problem is the following: How to measure the effectiveness of clustering user-generated definitions into different senses?

The primary issue lies in the lack of an evaluation framework for the clustering when the number of clusters is not known in advance. The second one is the lack of existing datasets with labelled instances that could be used for competing with existing work. Both these limitations of the state-of-the-art lead us to build a Ground Truth dataset for evaluation, and then to develop an evaluation method, that relies on the measurement of approximate correlation [14].

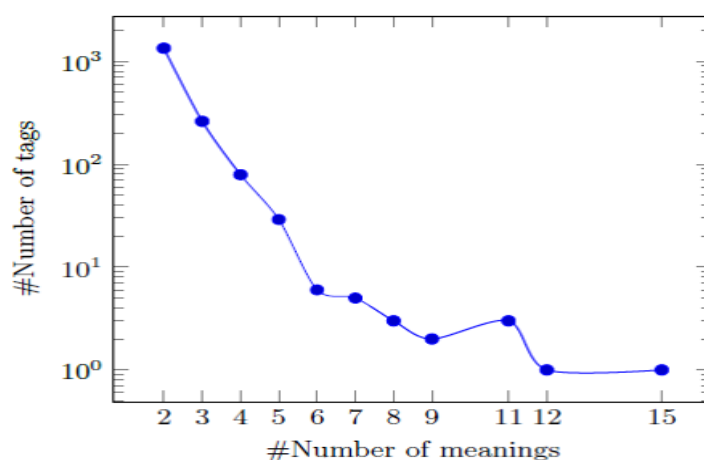


Fig. 4. Number of tags grouped by number of meanings.

4.3.1 Establishing Ground Truth:

We have built an ad hoc Web application, and participants have manually built the ground truth by clustering hashtags' definitions into meanings. The number of definitions in the folksionary and the Ground Truth is the same, yet ordered differently. The number of meanings is chosen independently by each participant. The web application eases enormously the manual work. To make a manual clustering, users group definitions that share the same meaning by adding a new meaning and sliding similar definitions on the same meaning.

4.3.2 Pairwise evaluation protocol:

We want to evaluate the \mathcal{E} function. In the following, we use the following notation:

- \mathcal{E}_{DGT} , the Dataset Ground Truth partitioning,
- \mathcal{E}_{DP} , the Dataset Prediction partitioning generated by our approach for the same dataset.

The evaluation objective is to measure how ϵ_{DP} performs towards ϵ_{DGT} . We are using a pairwise evaluation for this. For all pairs of definitions ($d_1; d_2$) for a word w , we define the following observations:

- if d_1 and d_2 are in the same cluster both in the ground truth and in the prediction, it is a true positive (TP),
- if d_1 and d_2 are in different clusters both in the ground truth and in the prediction, it is a true negative (TN),
- if d_1 and d_2 are in the same cluster in the ground truth, but in different clusters in the prediction, it is a false negative (FN),
- if d_1 and d_2 are in different clusters in the ground truth but in the same cluster in the prediction, it is a false positive (FP).

A synoptic view on this process is as follows :

- 1) For each word $w \in W$, enumerate all the pairs of user-generated definitions ($d_i; d_j$) in $(w) \times D(w)$ such as $i < j$,
- 2) Retrieve $C(s_i) = \epsilon_{DGT}(d_i)$ and $C(s_j) = \epsilon_{DGT}(d_j)$,
- 3) Retrieve $C(s'_i) = \epsilon_{DP}(d_i)$ and $C(s'_j) = \epsilon_{DP}(d_j)$,
- 4) Make observations (*TP*, *TN*, *FP*, or *FN*) depending on values of $C(s_i); C(s_j); C(s'_i); C(s'_j)$,
- 5) Compute a correlation measure for all words w . We have conducted observations on the entire dataset in order to measure the distance between the Ground Truth partitioning and the automatic partitioning generated by our approach. For this purpose we have performed a straightforward evaluation using a metric adapted to our dataset.

The most classical metrics one can find in the literature are the *F₁score* and the Matthews Correlation Coefficient (MCC). But these coefficients can be undefined when the denominator value is zero, which happens quite often in our case. The chosen metric is then Average Conditional Probability (ACP) [14] that smoothly takes into account such a case.

ACP is defined as follows if all the sums are non-zero:

$$ACP = \frac{1}{4} \left[\frac{|TP|}{|TP| + |FN|} + \frac{|TP|}{|TP| + |FP|} + \frac{|TN|}{|TN| + |FP|} + \frac{|TN|}{|TN| + |FN|} \right] \quad (4)$$

Otherwise, ACP is the average over those conditional probabilities that are defined.

4.3.3 Evaluation and interpretation

Our study intends to carry out comparisons across the performance of calculated measurements, in order to interpret the clustering output, and its proximity to the Ground Truth. As outlined above, we chose a graph-based algorithms MCL for our clustering approach because it can be used for detecting clusters from different shapes without specifying the clusters number in advance. The values of some parameters must be specified by the user as input, which remains a real challenge.

In this section, experimental results on our Folksionary are presented, in order to generate the combination of parameters representing the best tuning for the algorithm. After this tuning, we conduct assessments to measure the quality of our clustering approach compared to the Ground Truth. To achieve the first objective, several values of gamma-Exp (inflation exponent for Gamma operator), maxResidual (maximum difference between row elements and row square sum, measure of idempotence), and maxZero (maximum value considered zero for pruning operations) were tested (c.f. 3.2.3).

We carried out our experiments for the range of the following values : maxZero (10^{-1} ; 10^{-2} ; 10^{-3} ; 10^{-4} ; 10^{-5} ; 10^{-6} ; 10^{-7}), maxResidual (1; 0; 10^{-1} ; 10^{-2} ; 10^{-3}), gammaExp(1.4, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20). For each test, maxZero value is set and gammaExp value is varied with the maxResidual value in order to establish optimal values as said previously. Results substantially confirm that a good clustering requires a correct choice of parameters. The analysis clearly shows that ACP value keeps constant at **53.2%** for maxResidual = 1 and does not exceed **55.9%** for maxResidual = 0 regardless maxzero tested values. We also note that, for these maxResidual values 10^{-1} ; 10^{-2} and 10^{-3} , the ACP value converges rapidly to very good values for small values of gammaExp while decreasing maxZero. For example with maxZero= 10^{-1} ACP remains constant at **89.2%** starting from gammaExp =6 and begins to increase at 8, 10, 14, 18, 20 for the values of maxZero 10^{-2} ; 10^{-3} ; 10^{-4} ; 10^{-5} ; 10^{-6} ; 10^{-7} respectively. Results reported on Figure 4 represent the % of ACP by variant maxZero and setting maxResidual to 10^{-3} , and Figure 5 represent the % of ACP by varying maxResidual and setting maxZero to 10^{-1} .

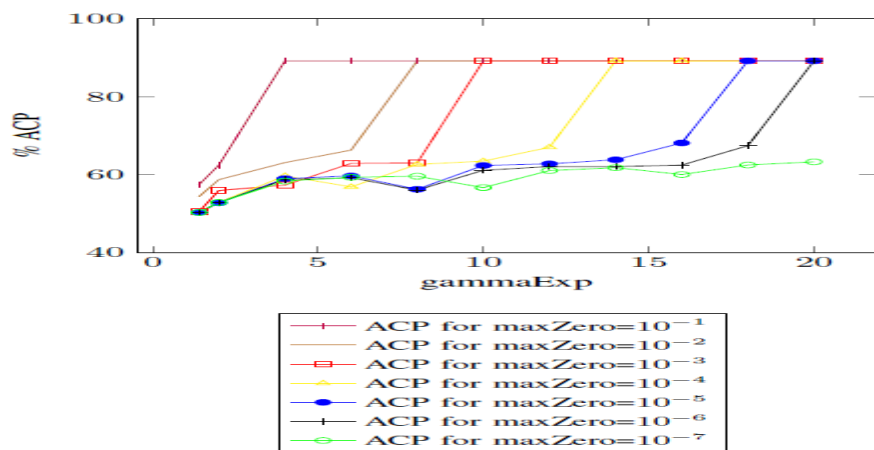


Fig5.% ACP by variant maxZero and setting maxResidual

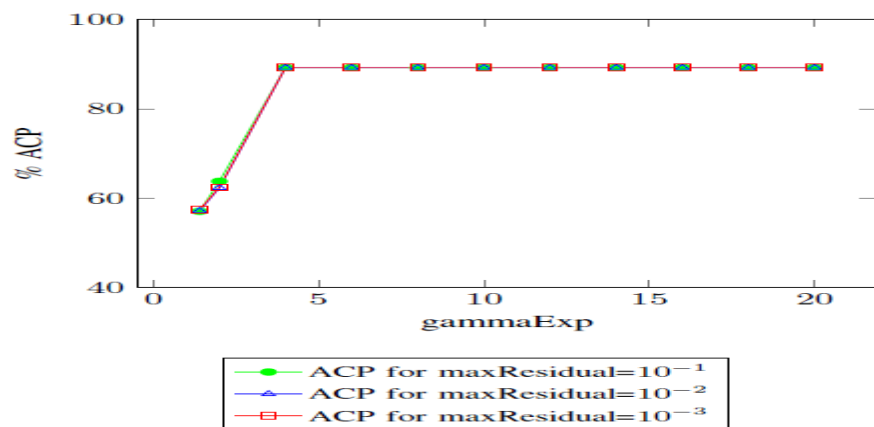


Fig6. % ACP by variant maxZero and setting maxResidual

We conclude that the best combination of MCL parameters for our dataset is maxZero = 0.1 and maxResidual value in the interval [10^{-1} ; 10^{-3}], while gammaExp value can begin from 6. Results reported on Table I represent the ACP analysis for maxZero = 10^{-1} .

Table1.The Acp Analysis For Maxzero = 10^{-1}

gammaExp	maxZero= 10-1				
	r=1	r=0	r=10-1	r=10-2	r=10-3
20	53.20	50.98	89.21	89.21	89.21
18	53.20	50.24	89.21	89.21	89.21
16	53.19	50.80	89.21	89.21	89.21
14	53.20	50.09	89.21	89.21	89.21

12	53.20	51.51	89.21	89.21	89.21
10	53.20	49.43	89.21	89.21	89.21
8	53.19	48.05	89.21	89.21	89.21
6	53.20	51.20	89.21	89.21	89.21
4	53.20	60.47	63.78	62.30	62.44
2	52.95	49.94	89.21	89.21	89.21
1.4	53.20	54.79	57.07	57.18	57.41

As shown in table I ACP value keeps constant at 89.2% starting from gammaExp =6 and in the interval $[10^{-1}, 10^{-3}]$ of maxResidual. Then to choose the best value for both parameters, we based on another criterion which is the temporal complexity. We opted for the combination which converges faster than the others. Table II summarizes the combinations for the different tests and their execution time.

Table2.The Acp Analysis For Maxzero = 10-1

gammaExp	maxZero= 10-1		
	r=10-1	r=10-2	r=10-3
20	34 min 7 sec	32 min 6 sec	31 min 40 sec
6	33 min 0 sec	31 min 23 sec	30 min 53 sec

We have pushed the value of gammaExp to 200 and 2000 and we noticed that the more its value grows, the more the execution time decreases. Then the best configuration of the MCL algorithm for us is : **maxZero= 10^{-1} , maxResidual= 10^{-3} and gammaExp=20.**

As a conclusion, from the experimental analysis carried out we see that results generated by the Automatic Partitioning with the best tuning of the MCL algorithm are close to those derived from Ground Truth with ACP=89,2%, which proves that our approach for definition-sense clustering achieved good results. Finally it should be noted that evaluating the performance of our clustering approach was not trivial, as the construction of manual Gound Truth is not an easy task, there is always a large variability in the number of clusters that humans generate for the same dataset. That is why, this dataset was enhanced by the confrontation between different manual partitioning made by different persons, so as to lower subjectivity and then have a good dataset for evaluation.

5. SEMANTIZED HASHTAGS FOR RECOMMENDATION IN E-LEARNING

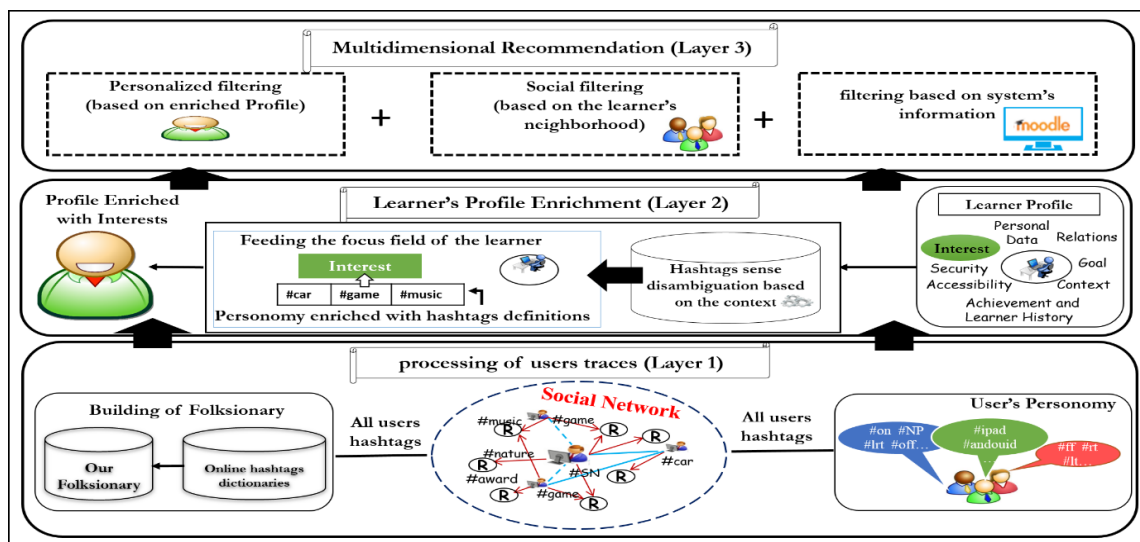


Fig7.Architecture for Multidimensional Recommendation on E-learning systems

Our question was how to make sense on the hahstags? where the idea of building a Folksonary (c.f. section~3). In this section the question is how to use Semantized hashtags for recommendation in e-learning. To activate this we presents our multidimensional recommendation architecture (figure 6), which intended to link the traces of users on social networks (hashtags) with their elearning profile and eventually make multidimensional recommendation of pedagogical ressourcecs.

5.1 Layer 1: Processing of Users Traces

Work in this layer is dedicated to ressemblance and treatment of learners writing on social networks, it's a preparatory step to make usable hashtags for other layers. On the one hand we collect hashtags for each learner separately for the construction of its personomie, on the other hand we crawl the most used hashtags on social network and their definitions from online hashtags dictionaries. Hashtags are then consolidated in a dictionary where they are disambiguated. The both results of this phase: the personomie gathering hashtags contained in learner's writings and representing its interests. And the folksonomy (c.f. 3) which consists in a dictionary that for each hashtag clusters its definitions in meanings.

5.2 Layer 2: Learner's Profile Enrichment

In this a learner's profile defined according to LMS-LIP standard. We carry out the enrichment of this profile with interests deduced from the personomie. In the previous step we only generate the personomie, in this layer we make treatment and we do a treatment on hashtags to disambiguate them according to their context of use in the tweet. The Lesk [31] algorithm is used to disambiguate hashtags and assign them best definition adapted to the context of use. Thus produced definition represent the interests of a learner as the hashtags are suitable to provide interesting information on the interests of users of social networks. We mention that to achieve disambiguation of personomie, pretreatment was realse on hashtags to disambiguate them and generate the senses candidates (c.f. 3).

We also note that the enriched profile can be further extended through the analysis of interest already deduced from the personomie. Based on the fact that learner's interests are the set of each hashtag's definitions, and hashtags can have relations with other ones. Discovering those relation allows to generate generate a new interests not explicitly mentioned outset in the personomie, where the hierarchical clustering proposed in the section above.

5.3 Layer 3: Multidimensional Recommendation

The main objective of our research is to improve the recommendation in e-learning environment, by proposing an approach that integrates new dimensions (social, annotations, traces, relations, profiles, etc.) in the recommendation process. Therefore we propose a recommendation model based at the same time on the enriched profile and other relevant social information such as learner's neighbors, or even other important informations released directly from the e-learning environment. In our overview we combine three types of filtering: personalized filtering based profile, social filtering and filtering based systems statistics. A synoptic view on this process is as follows:

Input: (Users Enriched Profile, Pedagogical resources)

➤ **profile based filtering**

- Compute similarity between the learner's interests and the description of pedagogical resources on the e-learning platform.
- Select relevant pedagogical resources based on the learner's interests.

➤ **Social Filtering**

- For each learner, compute the similarity of his/her profile and their neighbors profiles.
- Select the n most similar profiles to the target profile of the recommendation.
- Recommend the user with resources consulted or best evaluated by their closest neighbors.

➤ **Filtering based on systems statistics**

- Most visited resources by learners
- Among resources relevant to a learner, sort and select the top rated ones.
- Provide supplementary resources based on the history of the learner on the system.

Output: Recommended Resources

We focus on personalized recommendation based enriched profile within the e learning platform, although we chose Moodle as a case study. So for personalized recommendation we try to combine learners's interests to proposed courses according to their description.

6. CONCLUSION AND PERSPECTIVES

In this paper we have introduced the concept of Folksionary which consists in a dictionary that clusters each hashtag's definitions in meanings. We have also defined a four-steps process to build a folksionary. First we gather all definitions by crawling online services, we then apply a semantic distance measure between definitions for each hashtag. We perform a clustering that groups similar definitions into distinct meanings clusters. Clusters are finally presented under the form of a human-readable folksionary. We have conducted a validation of this process: we have developed a web application to build the Ground Truth where participants cluster the definitions. A pairwise evaluation of the results of our clustering process in comparison with the Ground Truth has been conducted. The Evaluation results show that our approach works not only in theory but also in practice: it performs well and produces good results for definition-sense clustering, by approaching Ground Truth with 89.2%. The very close next step concerns the development of techniques to discover other semantic relationships between tags: synonymy, hyperonymy, or part-of. In the long term our goal is to learn an ontology from the folksionary.

REFERENCES

- [1] LARA, Ruben, OLMEDILLA, Daniel, ARROYO, Sinuhe, et al. A semantic web services framework for distributed e-learning environments. Informe técnico, L3S Research Center, 2004.
- [2] CHUGHTAI, Muhammad Waseem, SELAMAT, Ali, GHANI, Imran, et al. E-learning recommender systems based on goal-based hybrid filtering. *International Journal of Distributed Sensor Networks*, 2014, vol. 2014.
- [3] ZAPATA, Alfredo, MENÉNDEZ, Víctor H., PRIETO, Manuel E., et al. Evaluation and selection of group recommendation strategies for collaborative searching of learning objects. *International Journal of Human-Computer Studies*, 2015, vol. 76, p. 22-39.
- [4] CHEN, Chih-Ming, LEE, Hahn-Ming, et CHEN, Ya-Hui. Personalized e-learning system using item response theory. *Computers & Education*, 2005, vol. 44, no 3, p. 237-255.
- [5] KLAŠNJA-MILIĆEVIĆ, Aleksandra, VESIN, Boban, IVANOVIĆ, Mirjana, et al. E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education*, 2011, vol. 56, no 3, p. 885-899.
- [6] FLORENCE SÈDES, Mihaela Brut. Modélisation basée sur des ontologies pour développer des recommandations personnalisées dans les systèmes hypermédia adaptatifs.
- [7] BERKANI, Lamia, CHIKH, Azeddine, et NOUALI, Omar. Recommandation personnalisée des ressources dans une communauté de pratique de e-learning. Une approche à base de filtrage hybride. In : *INFORSID*. 2013. p. 131-138.
- [8] ZOUAG, Rafika. Filtrage collaboratif des objets pédagogiques. 2014. Thèse de doctorat.
- [9] KHRIBI, Mohamed Koutheaïr, JEMNI, Mohamed, et NASRAOUI, Olfa. Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. In : *Advanced Learning Technologies*, 2008. *ICALT'08*. Eighth IEEE International Conference on. IEEE, 2008. p. 241-245.
- [10] IMS GLOBAL LEARNING CONSORTIUM, Inc, et al. IMS learning design information model. *IMS-LD*), version, 2003, vol. 1.
- [11] LESK, Michael. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In : *Proceedings of the 5th annual international conference on Systems documentation*. ACM, 1986. p. 24-26.
- [12] GALE, William A., CHURCH, Kenneth W., et YAROWSKY, David. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 1992, vol. 26, no 5-6, p. 415-439.
- [13] YAROWSKY, David. Unsupervised word sense disambiguation rivaling supervised methods. In : *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1995. p. 189-196.
- [14] IDE, Nancy et VÉRONIS, Jean. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics*, 1998, vol. 24, no 1, p. 2-40.

- [15] SMALL, Steven Lawrence. Word expert parsing: A theory of distributed word-based natural language understanding. 1980.
- [16] DAHLGREN, Kathleen. Naive semantics for natural language understanding. Dordrecht : Kluwer, 1988.
- [17] LESK, Michael. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In : Proceedings of the 5th annual international conference on Systems documentation. ACM, 1986. p. 24-26.
- [18] GHENNAME, Merieme, SUBERCAZE, Julien, GRAVIER, Christophe, et al. A hashtags dictionary from crowdsourced definitions. In : Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on. IEEE, 2014. p. 39-44.
- [19] BANERJEE, Satanjeev et PEDERSEN, Ted. An adapted Lesk algorithm for word sense disambiguation using WordNet. In : Computational linguistics and intelligent text processing. Springer Berlin Heidelberg, 2002. p. 136-145.
- [20] ENRIGHT, Anton J., VAN DONGEN, Stijn, et OUZOUNIS, Christos A. An efficient algorithm for large-scale detection of protein families. Nucleic acids research, 2002, vol. 30, no 7, p. 1575-1584.
- [21] BROHEE, Sylvain et VAN HELDEN, Jacques. Evaluation of clustering algorithms for protein-protein interaction networks. BMC bioinformatics, 2006, vol. 7, no 1, p. 488.
- [22] BRUYNOOGHE, M. Classification ascendante hiérarchique des grands ensembles de données: un algorithme rapide fondé sur la construction des voisinages réductibles. Cahiers de l'Analyse des Données, 1978, vol. 3, no 1, p. 7-33.
- [23] OUBAHSSI, Lahcen. Conception de plates-formes logicielles pour la formation à distance, présentant des propriétés d'adaptabilité à différentes catégories d'utilisateurs et d'interopérabilité avec d'autres environnements logiciels. 2005. Thèse de doctorat. Paris 5

AUTHORS' BIOGRAPHY



Merieme Ghennane^{1,2} received her master's Mohammed V Souissi degree in computer Engineering from University Mohammed V de Rabat, 2011. Her research interest contains Social Network, Semantics Web and E-learning. She is currently the Ph.D. student of Social and Semantics Web for E-learning Systems at Hubert Curien, University Jean Monnet and LeRMA University Mohammed V de Rabat.



Mounia Abik¹ I received a PhD from ENSIAS de Rabat in 2009. My main research interests focus on e- Learning, Mobile and Pervasive Learning, Semantic Web, Cloud Learning. My interests today focus Knowledge extraction in the connected learning environments, Semantic Web. I'm currently Professor Ability in ENSIAS-Rabat.



Julien Subercaze² I received a PhD from INSA de Lyon in 2010, after a CS Master degree from KIT (Karlsruhe), and a CS bachelor degree from Université de Rennes. She is currently the Ph.D. I started working on semantic agent during my PhD. My main research interests today focus on Data on the Web, Scalable Algorithms, Web information extraction, Social and Semantic Web, and their applications for Cloud Computing, Big Data, and Multimedia.



Christophe Gravier² I received a PhD from Université Saint-Etienne in 2007, after a CS Master degree from INSA Lyon, and a French engineer diploma from Télécom Saint-Etienne. Since approximately 2010, my research has somehow shifted away from online engineering. My main research interests today at SATIN research team focus on Web information extraction, Social and Semantic Web, and their applications for Cloud Computing, Big Data, and Multimedia.



Frederique Laforest² I got an engineer master in Computer Science from INSA de Lyon in 1993. I received a PhD from INSA de Lyon in 1997 and an Habilitation to Drive Research (HDR) from Université Lyon I in 2007. I was Associate Professor in INSA Lyon and in the LIRIS lab. from 1998 to 2011. I got the current Professor position in Université Jean Monnet in septembre 2011. I head the Satin research team of the Hubert Curien lab.



Rachida Ajhoun¹ I received a PhD from Ecole Mohammadia d'ingénieurs de Rabat Morocco in 2001, I'm Professor position in ENSIAS – Université Mohammed V de Rabat and Director of e-Learning Center. My main research interests today focus on Pervasive Learning, Collaboration, Context-aware, Access Control, Accessibility, and Enterprise 2.0. I head the LeRMA research team of the ENSIAS.