An Integrated Approach for Context-Aware Query Recommendation in Folksonomies

Chiraz Trabelsi and Sadok Ben Yahia

Université de Tunis El Manar, Faculty of Sciences of Tunis, Laboratory LIPAH, 1060 Tunis, Tunisia

RÉSUMÉ. *L'essor des sites collaboratifs sur Internet a permis la naissance de nouvelles formes d'indexations des contenus du Web, créées librement par les usagers et partagées au sein de réseaux sociaux, baptisées sous le nom de folksonomie. Considérées comme source de données, ces dernières s'avèrent d'un grand intérêt pour la Recherche d'Information. Cependant, la démarche de recherche dans les folksonomies diffère des stratégies de recherche de la traditionnelle médiation des moteurs de recherche dans la mesure où elle ne prend pas en considération l'aspect social et comportemental des usagers. Ainsi, afin d'assister les usagers et de leur permettre l'accès le plus pertinent à l'information, nous proposons une approche basée sur le couplage des Modèles de Markov Cachés (MMC) et des concepts triadiques pour la prédiction des intentions de recherches dans les folksonomies. Les résultats obtenus sur une folksonomie réelle s'avèrent prometteurs et ouvrent de nombreuses perspectives*

ABSTRACT. *Collaborative tagging systems have recently emerged as one of the most popular tools for Web search users to find their desired information. Indeed, the informal social classification structure in these systems, also known as folksonomy, provides a convenient way to annotate resources by allowing users to use any keyword or tag that they find relevant. In turn, the flat and non-hierarchical structure with unsupervised vocabularies leads to low search precision and poor resource navigation and retrieval. The widely keyword-based approaches used for locating information on the Web, are hence not straightforwardly adaptable to folksonomies. This drawback has created the need for an effective framework to support folksonomy users in effectively retrieving the resources matching their real search intents. The primary focus of this paper is to propose an integrated approach for context-aware query recommendation in folksonomies that exploits the power of both Hidden Markov Models (HMMs) and triadic concepts. We demonstrate through carried out experiments that the proposed approach yields mainly highly Precise and highly Relevant tag query recommendation*

MOTS-CLÉS : *MMC, folksonomie, intention de recherche, contexte, concept triadique*

KEYWORDS: *HMMs, folksonomy, search intent, query context, triadic concept*

1. Context and related work

Search and retrieval are vital parts of content categorization of the Web 2.0 age, and are increasingly receiving attention with the growing use of collaborative tagging systems and the explosion of sharing contents on the Web. Recently, collaborative or social tagging sites such as DEL.ICIO.US[1], BIBSONOMY [2] or FLICKR[3] have achieved widespread success on the Web due to their simplicity to browse and search an important body of shared resources. In these sites, users annotate resources such as Web pages, blog posts or pictures using a freely chosen set of keywords *aka* tags. The resulting structures are called folksonomies[4], that is, "taxonomies" created by the "folk". Considered as a tripartite hyper-graph (Mika) of tags, users and resources, folksonomies have become one of the most popular tools for Web users to find their desired information (Pan, Taylor and Thomas). However, one significant problem arising in folksonomies search stands in the tag ambiguity : tags that have several meanings, *e.g.,* "Java" as coffee or a programming language or an island in Indonesia. Moreover, folksonomy tags are unstructured as assigned tags to a given resource are simply enumerated in a list and finally, no special organization or categorization of the tags is made (by the folksonomy site). As per Golder and Huberman (Golder and Huberman), the main problems of social tagging systems include *ambiguity, lack of synonymy and discrepancies in granularity*. Specifically, when a user tries to retrieve resources using a certain tag, *e.g.,* "mackintosh", he can receive restrictive results since the system retrieves resources tagged with that particular tag regardless of the eventual vocabulary derivations or synonyms of this later, *e.g.,* "mac", "macintosh" or "mack". Indeed, there is no "standard" or "optimal" way to issue queries to a folksonomy search engine, and it is well recognized that query formulation is a bottleneck issue in the usability of search engines. Query recommendation is thus a promising direction for improving the usability of folksonomy search engines (Bouadjenek, Hacid, Bouzeghoub and Daigremont). The explicit task of query recommendation is to help users formulate queries that better represent their search intent during folksonomy search interactions.

In addition, as it has been forcefully argued that exploiting search contexts can improve information retrieval systems, introducing tag query search contexts can help better understand users' search intents and thus improve query prediction accuracy. In recent years, some researchers realized the importance of search context. In (He, Jiang, Liao, Hoi, Chang, Lim and Li) and (Cao, Jiang, Pei, Chen and Li), two context-aware approaches to query recommendation were proposed. Cao et *al.,* (Cao, Jiang, Pei, Chen and Li) proposed a general context-aware model for query suggestion and ranking. Whereas, in (He, Jiang, Liao, Hoi, Chang, Lim and Li), the authors proposed a novel sequential query prediction approach, based on Hidden Markov Models (HMM) training, that tries to grasp a user's search intent based on his past query sequence mined from massive search engine logs. These works confirmed that query

1. http ://www.delicious.com
2. http ://www.bibsonomy.org
3. http ://www.flickr.com
4. http ://wwww.vanderwal.net/folksonomy.html

search contexts are effective for disambiguating Web queries and can help improve the quality of multiple search services. However, because these aforementioned approaches ignored the three dimensional relationship among users, resources, and queries, the users tagging behaviors was not accurately profiled, and thus the suggestion quality based on the query and the resource data is not satisfactory. Indeed, regardless of their inadaptability to folksonomies search, our work has one fundamental difference from theses previous HMMs session-based approaches since folksonomies provide a three-dimensional dataset (users, queries and resources) instead of a usual two-dimensional web log data (queries and resources). Although, in (Lerman, Plangprasopchok and Wong), a HMM based approach is proposed in order to represent the users behavior of the FLICKR collaborative tagging system, the authors have only tackled a particular form of a folksonomy, *i.e.,* FLICKR, where a resource is only tagged by its owner, limiting its applicability to other contexts. More recently, Bouadjenek et al., (Bouadjenek, Hacid, Bouzeghoub and Daigremont), proposed an approach for social and personalized query expansion aiming to transform an initial query Q to another query Q' enriched with close terms that are mostly used by a given user and his/her social relatives. For doing that, the authors proposed to split the folksonomy into three bipartite graphs, *i.e.,* User-Tag, User-Document, and Tag-Document. Even though interesting and simple, this approach presents some drawbacks since the shift of dimension will inevitably lead to a loss of information.

In this paper, we propose an integrated approach for context-aware query recommendation in folksonomies that straightforwardly handle the triadic form of a folksonomy in order to keep track of its different elements : users, tags and resources. We intend to address the following challenges : 1) How to model and mine users search intents taking into account the specific three-dimensional structure of folksonomies ? ; and 2) How to provide an effective and efficient query recommendation method that produces simultaneously high *Precision* and *Recall* in folksonomies search ?.

First, to the best of our knowledge, none of existing works on query prediction can provide a good fit for the implicit dependencies between users, resources, and tags as fundamental property of the folksonomy structure. In our approach, tag assignments information is considered as an important part of context. Although many existing works (G.Begelman, Keller and F.Smadja, Mika) using clustering or statistical similarity metrics techniques studied how to use the co-occurrence of tags to find groups of related tags that may represent the tag contexts, none of them handled the semantic relatedness concurrently embodied in the different frequencies of co-occurrences among users, resources and tags in the folksonomy. Hence, instead of mining patterns of individual tag queries regardless the correlation between users, tag queries and resources, we summarize tag queries into search intents represented as triadic concepts that stresses users-queries-resources correlations (Jäschke, Hotho, Schmitz, Ganter and G.Stumme).

Indeed, the triadic concept structure describes the correlation between three types of sets : *(i)* the set $\mathcal{T}_1$ of semantically related tags queries, representing the tag query context ; *(ii)* the set $\mathcal{ID}_1$ of the associated users, *i.e.*, users whose have tagged by $\mathcal{T}_1$

and ; *(iii)* the set of related resources $\mathcal{RS}_1$ (representing the resource context) *i.e.*, which were assigned with $\mathcal{T}_1$ by users $\mathcal{ID}_1$. Hence, triadic concepts allow grouping semantically related tags taking into account the users' tagging and\or searching behavior in a folksonomy. Indeed, in folksonomies, the usage of tags of users with similar interests tends to converge to a shared vocabulary. To this end, we use a previous proposed algorithm (Trabelsi, Jelassi and Yahia), called TRICONS, for a scalable mining of users search intents that are hidden in a folksonomy.

Second, the query recommendation phase is straightforward, and the key issue is which model to choose for the particular search intent prediction problem ?. We surveyed a wide range of statistical models and narrowed down our choice to the Hidden Markov models (HMMs for short). This is because HMMs are parametric approaches to accurately estimate state-transition probabilities, and have proven powerful in modeling complex sequences in many fields such as in document classification (N.Tsimboukakis and G.Tambouratzis), in speech recognition (Rabiner) and more recently in context-aware web search (Cao, Jiang, Pei, Chen and Li, He, Jiang, Liao, Hoi, Chang, Lim and Li). Indeed, for the problem of search intent prediction, we are only interested in predicting the next tag query a user is likely to ask rather than labeling or predicting an entire follow-up sequence of observations. Furthermore, for our current formulation of user search intent, we can not directly model tag queries or tag query contexts as states, and yet assume them to be generated from some hidden states.

To sum up, we introduce in this paper an integrated approach for context-aware query recommendation in folksonomies which is composed of three main stages as follows :

1) *Patterns extraction :* taking part offline and during which user's queries sequences and user search intent are extracted from the considered folksonomy,

2) *HMM learning :* that also takes part offline and during which the HMM is trained given the previously extracted patterns,

3) *Matching and prediction :* this stage takes part online and its goal is to recommend the queries that best fit the user's information need. Matching and prediction stage stands in the identification of the current user query context and then the prediction of his next query according to the HMM states.

The remainder of the paper is organized as follows. Section 2 is dedicated to describe the first stage of our approach, during which user's queries sequences and users' search intents are mined. We introduce in section 3, the second stage of our approach, during which the HMM training steps are handled. We describe later in Section the online matching and prediction stage, aiming at identifying the tag query context and then predict the next user query. We dedicate Section 5 for underpinning, through an illustrative example based on a sample taken from a real dataset, the guidelines of our approach. The experimental study of our approach is illustrated in Section 6. Section 7 concludes this paper and sketches avenues for future work.

2. Patterns extraction

The goal of this stage is to learn users' search behavior by identifying users' search intents behind queries. Although query log data has so far proven extremely useful for learning user behavior and improving search engine accuracy (Fonseca, Golgher, Péssas, Ribeiro-Neto and Ziviani, Jones, Rey, Madani and Greiner, Baeza-Yates and Tiberi), the availability of folksonomies search log data is unfortunately limited because of serious privacy concerns. Fortunately, tag data by nature is all publicly available, and the amount of data is increasing rapidly. Therefore, for conducting our approach we make use of the duality hypothesis of search and tagging, two important behaviors of web users (Benz, Hotho, Jäschke, Krause and Stumme, Bischoff, Firan, Nejdl and Paiu, Krause, Hotho and Stumme, Mei, Jiangy, Suz and Zhai).

DUALITY HYPOTHESIS : The users' bookmarking and tagging actions reflect their personal relevance judgement.

The duality hypothesis immediately suggests that the query log data and the tagging data can be equally valuable for inferring a user's information preferences and thus improving many information management tasks such as search and information recommendation (Mei, Jiangy, Suz and Zhai). Consider, for example, a user who assigned the tag "java" to the Apache Lucene homepage (http ://lucene.apache.org), then we assume that the user will consider this web page as relevant if he issues "java" as a query. Actually, the duality hypothesis have opened up a highly promising new directions for using tagging data to analyze user behavior and improve search accuracy (Heymann, Koutrika and Garcia-Molina, S.Xu, Bao, Fei, Su and Yu).

Defined as a set of assignments, *i.e.*, triples (resources, users, tags), folksonomy can be seen as the other side of the "medal", *i.e.*, the log files (Krause, Jäschke, Hotho and Stumme). As logdata contains queries, clicks and session IDs, the classical dimensions of a folksonomy can be reflected : Queries or query words represent tags, session IDs correspond to users identifiers, and the URLs clicked, *i.e.,* resources accessed, by users can be considered as the resources that they tagged with the query words. Therefore, we define the folksonomy as following :

Définition 1 (FOLKSONOMY)

A folksonomy is a set of tuples $\mathcal{F}^{\nabla} = (\mathcal{ID}, \mathcal{Q}, \mathcal{RS}, \mathcal{G})$*, where :*

– $\mathcal{ID}$*,* $\mathcal{Q}$*,* $\mathcal{RS}$ *respectively define, all user IDs, the submitted queries and the accessed resources,*

– $\mathcal{G} \subseteq \mathcal{ID} \times \mathcal{Q} \times \mathcal{RS}$ *represents a triadic relation, where each* $g \subseteq \mathcal{G}$ *can be represented by a triplet :* $g = \{(id, q, rs) \mid id \in \mathcal{ID}, q \in \mathcal{Q}, rs \in \mathcal{RS}\}$*. Roughly speaking, a user identified by* id*, retrieved the resource* rs *queried by the query* q*.*

Regarded as a tripartite graph of users identifiers, submitted queries and accessed resources, the folksonomy can be, formally, represented as a triadic context (Jäschke, Hotho, Schmitz, Ganter and G.Stumme).

Exemple 1 *Fig. 1 illustrates an example of a folksonomy $\mathcal{F}^{\nabla}$ where $\mathcal{ID}$ = {id_1, id_2, id_3, id_4}, $\mathcal{Q}$ = {q_1, q_2, q_3, q_4, q_5} and $\mathcal{RS}$ = {rs_1, rs_2, rs_3}.*

Note that each $\times$ is a triadic relationship between a user identifier belonging to $\mathcal{ID}$, a query from $\mathcal{Q}$ and the resource accessed belonging to $\mathcal{RS}$.

$\mathcal{ID}/\mathcal{RS}-\mathcal{Q}$	rs_1					rs_2					rs_3				
	q_1	q_2	q_3	q_4	q_5	q_1	q_2	q_3	q_4	q_5	q_1	q_2	q_3	q_4	q_5
id_1		$\times$	$\times$	$\times$			$\times$	$\times$	$\times$			$\times$	$\times$	$\times$	
id_2		$\times$	$\times$	$\times$		$\times$	$\times$	$\times$	$\times$		$\times$	$\times$	$\times$	$\times$	
id_3		$\times$	$\times$	$\times$		$\times$	$\times$	$\times$	$\times$		$\times$	$\times$	$\times$	$\times$	
id_4						$\times$			$\times$		$\times$			$\times$	

Tableau 1. *An example of a folksonomy $\mathcal{F}^{\nabla}$*

The patterns extraction stage proceeds concurrently by firstly retrieving user's queries sequences from a folksonomy and mine users' search intents. Thereafter, the results of the previously steps, will be used for the HMM training.

2.1. *Step 1 : User's queries sequences extraction :*

In this step we are interested in discovering the queries sequences SL_i of each user idu_i. Hence, we must firstly collect users sessions S_i from the folksonomy. Lets us, at first, give the definition of a user session.

$S_1 := \{\{q_{1,1}, q_{1,2}, q_{1,3}\}, rs_{1,1}\}; \{\{q_{1,1}, q_{1,4}\}, rs_{1,2}\}$
$S_2 := \{q_{2,3}, q_{2,4}, rs_{2,3}\}$
$S_3 := \{\{q_{3,2}, q_{3,3}\}, rs_{3,4}\}; \{q_{3,4}, q_{3,5}, rs_{3,6}\}$

Tableau 2. *Users' sessions example*

Définition 2 (USER SESSION)

A user session S_i, related to a user idu_i, is defined as :

$$S_i := \{\{\text{User queries } q_{S_i,p}\}, rs_{S_i,j}\}.$$

With, $rs_{S_i,j}$:= the resource j accessed by the user idu_i in the session S_i ;
and $q_{S_i,p}$:= the p ordered submitted query in S_i.

Table 2 illustrates an example of users sessions. For example, the user session S_2, highlights that the user idu_2 has retrieved the resource $rs_{2,3}$ after submitting the two queries $q_{2,3}$ and $q_{2,4}$.

Once the users sessions are collected, we derive user's queries sequences by keeping, for each user, the sequences of queries related to his session and discard useless

information. An example of user's queries sequences associated to the Table 2 is given in the following : SL_1 : (($q_{1,1} \Longrightarrow q_{1,2} \Longrightarrow q_{1,3}$) ; ($q_{1,1} \Longrightarrow q_{1,4}$)), SL_2 : ($q_{2,3} \Longrightarrow q_{2,4}$), SL_3 : (($q_{3,2} \Longrightarrow q_{3,3}$) ; ($q_{3,4} \Longrightarrow q_{3,5}$)) where SL_i, describes query sequences of the user idu_i.

2.2. *Step 2 : User search intent mining :*

The second step of the model-learning step is to mine users' search intents from the folksonomy.

Hence, since, at one hand, different users may submit different queries to describe the same search intent and on the other hand, different users sharing the same interest for a specific topic, may retrieve different resources even if they submit exactly the same query, therefore we define a user search intent in a folksonomy as the common interest shared by a community of users ID for a retrieved set of resources RS queried by a certain set of queries Q.

Consequently, a search intent can be, formally, represented, in a folksonomy $\mathcal{F}^{\nabla} = (\mathcal{ID}, \mathcal{Q}, \mathcal{RS}, \mathcal{G})$, as a triadic concept $\mathcal{IT} = (U', T', R')$ where $U' \subseteq \mathcal{ID}$, $T' \subseteq \mathcal{Q}$, and $R' \subseteq \mathcal{RS}$ with $U' \times T' \times R' \subseteq \mathcal{G}$. Indeed, mining tri-concepts to discover and model users' search intents, allows to address the sparseness of queries and interpret users' information needs more accurately. The users' search intents are therefore obtained by applying the TRICONS algorithm (Trabelsi, Jelassi and Yahia) on the folksonomy $\mathcal{F}^{\nabla}$. TRICONS takes as input the folksonomy $\mathcal{F}^{\nabla}$ as well as three user-defined thresholds : $id-minsupp$, $q-minsupp$ and $rs-minsupp$ and outputs the set of all frequent tri-concepts, *i.e.,* search intents, that fulfill these aforementioned thresholds. For example, the search intent IT_1 = {(id_1, id_3, id_4), (q_4, q_5),(rs_1, rs_2)} is obtained by applying TRICONS algorithm on the folksonomy depicted by Fig. 1, with $id - minsupp$ = $q-minsupp$ = $rs-minsupp$ = 2. Roughly speaking, the search intent IT_1, highlights that the community of users (id_1, id_3, id_4) share the same interest in the resources (rs_1, rs_2) queried by q_4 and q_5.

Given the user's queries sequences and the users' search intents, previously extracted, we proceed in the next section with the HMM training .

3. HMM **learning**

For the second stage of our approach, we are interested in training HMM. The HMM starts with a finite set of states. Transitions among these states are governed by a set of probabilities (*i.e.,*transition probabilities) associated with each state. Assuming that there are two types of states in a HMM : the observable states and the hidden ones (Rabiner), thereby, we define user's queries sequences as the observable states in the HMM, whereas the hidden states are modeled by the users' search intents. Note that, if we model individual queries and resources directly as states in the HMM, then we not only increase the number of states and thus the complexity of the model, but also lose the faithful preservation of the semantic relationship among the queries and the accessed resources within the same search intent. Therefore, given the set of hidden

states $S = \{s_1,\ldots,s_{ns}\}$, we denote the set of distinct queries as $\mathcal{Q}$ = $\{q_1,\ldots,q_{nq}\}$, the set of accessed resources $\mathcal{RS}s$ = $\{rs_1,\ldots,rs_{nrs}\}$ and a set of user $\mathcal{ID}$; $\mathcal{ID}us$ = $\{idu_1,\ldots,idu_{nidu}\}$, where ns is the number of states of the model, nq is the total number of queries, nrs is the total number of resources, $nidu$ is the number of users, and SL_i is a state sequence. Our HMM noted λ = (A, B, B', π), is a probabilistic model defined as follows :

– π = [$\ldots\pi_i\ldots$], the initial state probability, where π_i = P(s_i) is the probability that a state s_i occurs as the first element of a state sequence SL_i.

– B = [$\ldots b_j(q)\ldots$], the query emission probability distribution, where $b_j(q)$ = P(q | s_j), denotes the probability that a user, currently at a state s_j, submits a query q.

– B' = [$\ldots b_k(rs)\ldots$], the resource emission probability distribution, where $b_k(rs)$ = P(rs| s_k), denotes the probability that a user, currently at a state s_j, accesses the resource rs.

– A = [$\ldots a_{ij}\ldots$], the transition probability, where a_{ij} = P(s_j | s_i) that represents the transition probability from a state s_i to another one s_j.

Once the HMM is formalized, we proceed with learning its parameters (A, B, B', π) from a folksonomy. This is done by performing two distincts stages namely : *(i)* The initial HMM parameters values assignment ; and HMM parameters values re-estimation. In the following, we present each stage.

– **The initial** HMM **parameters values assignment :**

The goal of this stage is to compute the four sets of the HMM parameters : the initial state probabilities {P(s_i)}, the query emission probabilities {P(q_t| s)}, the resource emission probabilities {P(rs| s_k)}, and the transition probabilities {P(s_j | s_i)}.

- π_i = P(s_i) = $\frac{|\varphi(s_j)|}{|SL_c|}$ with :

  - SL_c = $\cup_{i\in 1,\ldots,t}\{E_i\}$ = total set of candidate states sequences to which could be matched a sequence of queries where E_i denotes the set of candidate states that could match a query from a given sequence of queries.

  - $\varphi(s_j)$= set of states sequences in SL_c starting from s_j.

- $b_j(q)$ = P(q| s_j) = $\frac{\sum_{rs\in\mathcal{RS}_j} Count(rs,q)}{\sum_{q\in\mathcal{Q}_j}\sum_{rs\in\mathcal{RS}_j} Count(rs,q)}$.

- $b_k(rs)$ = P(rs| s_k) = $\frac{\sum_{q\in\mathcal{Q}_k} Count(rs,q)}{\sum_{q\in\mathcal{Q}_k}\sum_{rs\in\mathcal{RS}_k} Count(rs,q)}$. where $Count(rs,q)$ = number of times the resource rs is accessed as an answer to the query q in the folksonomy.

- $a_{i,j}$ = P(s_j | s_i) = $\frac{CS(s_i,s_j)}{NC}$ with :

  - NC = the number of occurrences of s_j in SL_c.

  - $CS(s_i,s_j)$= the number of times the state s_i is followed by the state s_j in SL_c.

For more accurate predictions results, we are interested, in the next stage, in re-estimating the HMM initial values parameters to to yield a more accurate predictions.

– **The** HMM **parameters values re-estimation :**

We mainly give a brief presentation of the used algorithm namely, the Baum-Welch algorithm (Rabiner). In fact, Baum-Welch algorithm aims to find the maximum-likelihood estimate of the parameters of a HMM given a set of observed feature vectors.

Hence, considering the initial set $\lambda = (A, B, B', \pi)$. The Baum-Welch algorithm updates the parameters of λ iteratively until convergence. Thus, given the total set of observations *O*, *i.e.,* the set of all query sequences, the Baum-Welch algorithm finds :

λ^* **=** $\boldsymbol{argmax}_\lambda$**ln P(***O*$|\lambda$**)** - that is, the HMM λ^*, that maximizes the probability of the observation *O*.

Based on the Baum-Welch algorithm, we get out λ^* that relieves the set of observation sequences with the highest probability of occurrence.

4. Matching and prédiction

Once the HMM-learning step is performed, we proceed with the online query recommendation step. The main goal of this step stands in identifying the tag query context and then predict the next user query according to the next HMM state. Indeed, when a user submits a query q, two consecutive stages are carried out : *(i)* Matching the current user query q to its corresponding context according to HMM states ; and then *(ii)* Predicting the next HMM state which represents the user's search intent. Hence, the prediction process starts by looking for the most likely HMM state s_{MS} to which q could better belong. This is done by computing, for each HMM state, the value of the quantity $Mat_i = \pi_i \times b_i(q)$, where π_i is the initial probability of the state s_i and $b_i(q)$ is the emission probability of q at s_i. Therefore, the state with the highest value, *i.e.,* s_{MS}, of Mat_i will define the context of q.

Thereafter, the prediction of the user's search intent is then performed, by looking for the next state $s_{Next\texttt{MS}}$ of s_{MS}. This is obtained by computing the index value $Next\texttt{MS}$ as follows : $Next\texttt{MS} = argmax_j\{a_{\{MS,j\}} \times b_j(q)\}$, where q denotes a query belonging to the state s_j, successor of s_{MS} in the HMM.

Thus, the state $s_{Next\texttt{MS}}$ represents the most probably search intent to which the user may transit after submitting the query q. Queries belonging to the search intent represented by the state $s_{Next\texttt{MS}}$ will be suggested to the user in an increasing ranked list of probability. Likewise, the corresponding resources of the considered search intent could be recommended in the same way.

5. Illustrative example of the proposed approach on a real dataset

Fig. 1 represents a HMM with five states $\{s_1, s_2, s_3, s_4, s_5\}$ where each state denotes a user search intent, *i.e.*, It_1, It_2, It_3, It_4 and It_5, extracted by the algorithm TRICONS from a sample taken from the real test data collected from DEL.ICIO.US[5].

5. http ://www.delicious.com

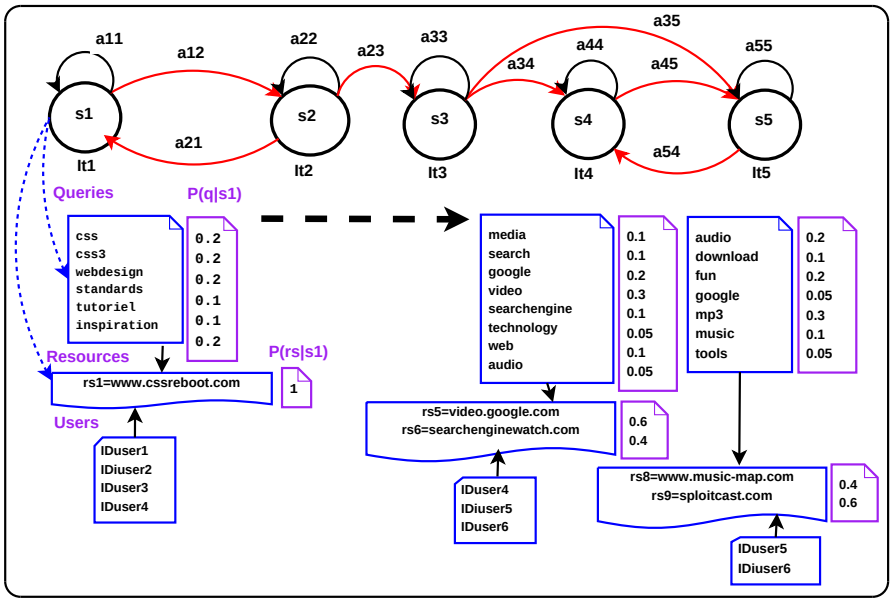

Figure 1. *An example of the proposed approach on a sample of a real dataset*

Each search intent is represented by a triplet, *i.e.*, the set of all queries frequently used by a set of users looking for a set of resources. The corresponding transition matrix A, and the distributions of the different probabilities of observation (of resources and queries) are obtained by computing probabilities as described in the previous sections. The corresponding HMM with five states is shown in Fig. 1. Suppose that the generated HMM with five states $\{s_1, s_2, s_3, s_4, s_5\}$ has a transition probability matrix as follows :

$$A = \begin{pmatrix} 0.4 & 0.6 & 0.0 & 0.0 & 0.0 \\ 0.2 & 0.5 & 0.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.3 & 0.2 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.3 & 0.7 \\ 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

And let us assume that $\pi = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$. Hence, considering the search intent represented by the state s_1, users have a probability of 0.4 to keep the same search intent and a probability of 0.6 to skip for a new search intent represented by the state s_2. For example, if a user submits the query *"audio"*, then the prediction process starts by looking for the most likely HMM state to which the query *"audio"* could better belong. This is obtained by computing for each of the five states, the quantity $Mat_i = \pi_i \times b_i(audio)$ including :

$Mat_1 = \pi_1 \times b_1(audio) = 0.2 \times 0 = 0$; $Mat_2 = Mat_3 = 0$; $Mat_4 = \pi_4 \times b_4(audio) = 0.2 \times 0.05 = 0.01$ and $Mat_5 = \pi_5 \times b_5(audio) = 0.2 \times 0.2 =$ **0. 04**.

Consequently, s_5 is the state which has the highest probability to represent the user's search intent for the query *"audio"*. Thus, the candidate resources (*rs8 :www.music-map.com/*) and (*rs9 :www.sploitcast.com/*), with the respective probabilities 0.4 and 0.6, are recommended to the user.

Furthermore, possible states transitions from s_5 are either s_4 or s_5 (*i.e*, a user may keep the same search intent). Thus, the corresponding candidate queries to be predicted, after the *"audio"*'s query submission, are computed by the following formula, $argmax_j\{a_{5,j} \times b_j(q)\}$ with $j \in \{4, 5\}$ (*i.e*, possible state transition from s_5) and q is a query belonging to the search intents represented by s_4 or s_5 states.

Otherwise, given that in the one hand : $Max(b_5(q)) = 0.3$ and $Max(b_4(q)) = 0.3$ for all queries q in the fifth and the fourth state respectively, and on the other hand $argmax_j\{a_{5,5} \times b_5(q), a_{5,4} \times b_4(q)\} = argmax_j\{0.12, 0.36\} = 4$, then the search intent to be predicted is represented by the state of index 4 (*i.e*, s_4). Thus, queries {*"video", "media", "google"*,... } belonging to the search intent represented by s_4 will be suggested to the user in an increasing ranked list of probability. Likewise, the corresponding resources of the considered search intent could be recommended in the same way.

6. Experimental evaluation

The evaluation of all folksonomy's query recommendation systems is still an open challenge. In fact, as an evidence of the lack of social bookmarking systems that exploit search intents prediction, as far as we know, there is no work with topic published via the scholarly literature. Hence, the evaluation of our approach is a complex task. In order to analyze the accuracy of our approach we adopted the common evaluation measures, namely Precision and Recall.

We carried out experiments on a dataset collected from a real-world social bookmarking system, *i.e*., DEL.ICIO.US[6]. The related folksonomy[7] $\mathcal{F}^{\nabla}$ contains 99 989 triples sessions, *i.e.,* tag assignments, 18 066 queries, *i.e.,* tags, 53 397 resources, *i.e.,* considered as accessed resources and 43 419 users.

6.1. *Baselines Models*

To the best of our knowledge, search intent prediction (using hidden markov models) in such social bookmarking systems have never been modeled before. Thus, for enhancing the effectiveness of our approach we have selected two baselines models for query prediction. The most popular queries recommender which predicts queries according to their global occurrence in the training data. On the other hand, the most

6. http ://www.delicious.com
7. Around 10 MB in size (compressed) and is freely downloadable http ://data.dai-labor.de/corpus/delicious/

popular query aware recommender, which ranks queries according to their global co-occurrence in the training set, with the query tag in the test set.

For each of the algorithms of our evaluation, we briefly describe in the following the specific settings used to run them.

– **Most popular queries recommender** : For each query tag we counted in how many user sessions it occurs and used the top queries (ranked by occurrence count) as recommendations.

– **Most popular query aware recommender** : These recommenders weights query tags by their co-occurrence with a given query. We then used the most co-occurrent tag queries as a suggestion.

6.2. *Effectiveness of our approach*

We assess the performance of the proposed approach on query prediction using a supervised learning method. From each dataset, *i.e.,* the original and the pre-processed one, we build two test sets. Specifically, we randomly split each dataset into two parts, a training part and a test part. The training parts are used to estimate the model while the test parts are used for the evaluation. Hence, for a given sequence of queries, the first n queries are used for generating predictions, whereas, the remaining part $\mathcal{Q}_T$ of the queries is considered as the set of queries actually formulated by the user, as the ground truth. The performance is then assessed by *precision* and *recall* at different ranks K.

Suppose that for a user query q_T, the proposed approach predicts a list of queries $\mathcal{Q}_R$, thus, the measures of *Recall* and *Precision* are given as follows :

$$\textit{Recall} = \frac{|\mathcal{Q}_R \bigcap \{\mathcal{Q}_T \backslash q_T\}|}{|\mathcal{Q}_T \backslash q_T|}, \qquad \textit{Precision} = \frac{|\mathcal{Q}_R \bigcap \{\mathcal{Q}_T \backslash q_T\}|}{|\mathcal{Q}_R|}$$

Furthermore, in order to investigate the effect of the lack of semantics of folksonomies on the performances of the proposed approach, we also conducted the pre-processing stage that we previously proposed in (Trabelsi, Jrad and Yahia) on the considered folksonomy. Roughly speaking, the goal of the pre-processing stage is to clean and semantically enrich the folksonomy's tags. Hence, for each tag, différent filtering steps are performed to retrieve a related set of cleaned and enriched tags (Trabelsi, Jrad and Yahia). It is important to notice that the pre-processed set of tags is at most as big as the original vocabulary, *i.e.,* $Q^p \leq Q$ et $G^p \leq G$ (triple sessions). We assume that a properly established pre-processed folksonomy will probably boost the effectiveness of our query recommendation approach in terms of *Recall* and *Precision*. Hence, after the pre-processing stage, the pre-processed folksonomy contained $|Q^p| = 18\,312$ tag queries and $|G^p| = 99\,123$ triple sessions, *i.e.,* tag assignments.

We report in the following the obtained results of the performances of the proposed approach on query recommendation, on the two folksonomies, *i.e.,* the original and the pre-processed ones, averaged over all user sessions and after 6 test runs.

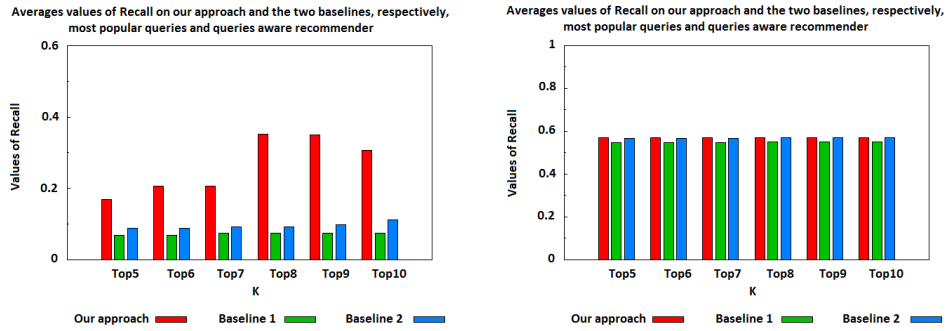

Figure 2. ***Left*** *: Averages of Recall on the original* DEL.ICI.OUS *dataset ;* ***Right*** *: Averages of Recall on the pre-processed* DEL.ICI.OUS *dataset.*

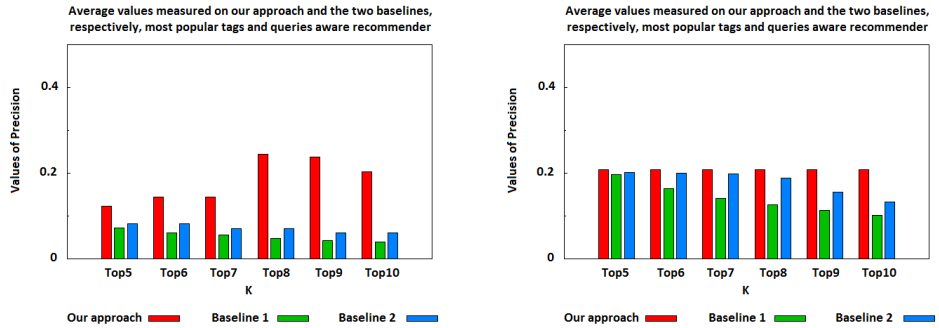

Figure 3. ***Left*** *: Averages of Precision on the original* DEL.ICI.OUS *dataset ;* ***Right*** *: Averages of Precision on the pre-processed* DEL.ICI.OUS *dataset.*

Figure 2 (Left), depicts averages of *recall* for different values of K, *i.e.,* the number of predicted queries, ranging from 5 to 10. Thus, according to the sketched histograms, we can point out that our approach outperforms the two baselines, for both, the original and the pre-processed datasets. In fact, as expected, the *Recall* values of the individual baselines are much lower than those achieved by our approach. Furthermore, the average *Recall* on the original folksonomy achieves high percentage for higher value of K. Indeed, for $K = 9$, the average *Recall* is equal to $0,351$, showing an increase of $51,85\%$ compared to the average *Recall* for $K = 5$. In this case, for a higher value of K, *i.e.,* $K = 9$, by matching current user's queries with their corresponding contexts, the proposed approach can produce all of the queries that are likely to be formulated by the user.

However, according to Figure 3 (Right), the percentage of *Precision* for the proposed model outperforms the two baselines over the two datasets. Our approach achieves the best results when we choose the value of K around 8. In fact, for $K = 5$, the mean precision, is equal to $12,3\%$. Whereas, for $K = 8$, it has an average of $24,5\%$ sho-

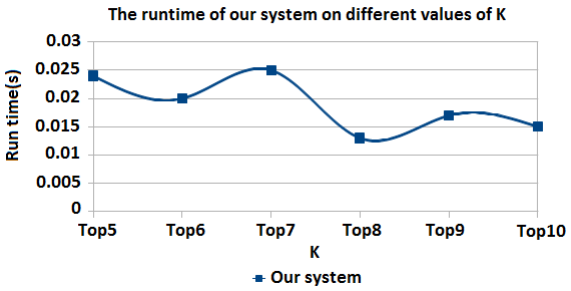

Figure 4. *The run time of our system online on the* DEL.ICI.OUS *dataset with different values of K*

wing a drop of the query prediction accuracy of $49,79\%$ *vs.* an exceeding about $19,6\%$ against the first baseline and around $17,4\%$ against the second one. These results highlight that the proposed approach can better improve query prediction accuracy even for a high number of predicted queries regardless the handled dataset. Moreover, our approach achieves a good coverage, since it produces predictions for 76% of queries contained in the test set $\mathcal{Q}_T$.

6.3. *Online evaluation*

We present in figure 4 the runtime of our system[8]. Since it is hard to measure the exact runtime of the model, we simulated an online execution of our system among the DEL.ICI.OUS dataset with different values of K, *i.e.,* the number of predicted queries, ranging from 5 to 10. Hence, for each tag query on the dataset, we report the average runtime of the related top K result. Figure 4 describes the online exécution of our system. In fact, the maximum value of run time is about $0.025(s)$, whereas the minimum value is around $0.013(s)$ which is efficient and satisfiable.

7. Conclusion

In this paper, we have introduced an integrated approach for context-aware query recommendation in folksonomies by using a powerful coupling model, based on an effective use of HMM and triadic concepts, to provide simultaneously high *Precision* and *Recall* in user's queries recommendation. We tackle the challenge of learning a large HMM from hundreds of thousands of user's sessions by summarizing individual queries, resources and users into search intents, formally, represented as triadic concepts which can greatly lower the number of the HMM states and allow to interpret users' information needs more accurately. Finally, we evaluated our proposed

8. The prediction system is implemented in *C++* (compiled with *GCC* 4.1.2) and we used an $IntelCorei7$ CPU system with 6 GB RAM. Tests were carried out on the Linux operating system Ubuntu 10.10.1.

approach on a large dataset, carried out from a real world folksonomy, through common metrics evaluation. We have also highlighted that tackling tag synonymy and polysemy problems boosts the query prediction relevance and accuracy.

Our future avenues for future work mainly address the focus on other more sophisticated Markov models such as variable length HMM in a folksonomy search. This includes modeling hidden states that represent users search intents, which could be an underlying semantic concept, especially with the help of domain knowledge such as the online ontologies. Indeed, the use of online ontologies may allow prediction systems to find out how specific the user interest is, and use this information to fine predictions. It remains to be seen whether more sophisticated models can further raise the performance bar for the query prediction in folksonomies. Moreover, DEL.ICIO.US now allows users to form links with other, *e.g.,* friends, groups. Such social links could be explored for further interest and search intent analysis.

8. Bibliographie

Baeza-Yates R., Tiberi A., « Extracting semantic relations from query logs », *Proceedings of the* 13^{th} *International Conference on Knowledge Discovery and Data mining*, ACM, New York, USA, p. 76-85, 2007.

Benz D., Hotho A., Jäschke R., Krause B., Stumme G., « Query Logs as Folksonomies », *In : Datenbank-Spektrum*, vol. 10, n° 1, p. 15-24, 2010.

Bischoff K., Firan C. S., Nejdl W., Paiu R., « Can all tags be used for search ? », *Proceedings of the* 17^{th} *ACM conference on Information and knowledge management, CIKM 2008*, ACM Press, Napa Valley, California, p. 193-202, October, 2008.

Bouadjenek M., Hacid H., Bouzeghoub M., Daigremont J., « Personalized social query expansion using social bookmarking systems », *Proceedings of the* 34^{th} *International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, ACM, New York, NY, USA, p. 1113-1114, 2011.

Cao H., Jiang D., Pei J., Chen E., Li H., « Towards context-aware search by learning a very large variable length hidden markov model from search logs », *Proceedings of the* 18^{th} *International Conference on World Wide Web, WWW 2009*, Madrid, Spain, p. 191-200, 2009.

Fonseca B., Golgher P., Péssas B., Ribeiro-Neto B., Ziviani N., « Concept-based interactive query expansion », *Proceedings of the* 14^{th} *ACM International Conference on Information and Knowledge Management*, Bremen, Germany, October, 2005.

G.Begelman, Keller P., F.Smadja, « Automated Tag Clustering : Improving Search and Exploration in the Tag Space », *Proceedings of the WWW Collaborative Web Tagging Workshop*, Edinburgh, Scotland, 2006.

Golder S., Huberman B., The structure of Collaborative Tagging Systems, Technical report, HP Labs, 2005.

He Q., Jiang D., Liao Z., Hoi S., Chang K., Lim E., Li H., « Web Query Recommendation via Sequential Query Prediction », *Proceedings of the 2009 IEEE International Conference on Data Engineering*, IEEE Computer Society, Washington, USA, p. 1443-1454, 2009.

Heymann P., Koutrika G., Garcia-Molina H., « Can Social Bookmarking Improve Web Search ? », *Proceedings of the First ACM International Conference on Web Search and Data Mining*, ACM, Stanford University, USA, February, 2008.

Jäschke R., Hotho A., Schmitz C., Ganter B., G.Stumme, « Discovering shared conceptualizations in folksonomies », *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 6, p. 38-53, 2008.

Jones R., Rey B., Madani O., Greiner W., « Generating query substitutions », *Proceedings of the* 15^{th} *International Conference on World Wide Web*, Edinburgh, Scotland, p. 387-396, 2006.

Krause B., Hotho A., Stumme G., « A Comparison of Social Bookmarking with Traditional Search », *Proceedings of the* 30^{th} *European Conference on IR Research, Advances in Information Retrieval, ECIR 2008*, Springer, Glasgow, Scotland, p. 101-113, 2008a.

Krause B., Jäschke R., Hotho A., Stumme G., « Logsonomy - Social Information Retrieval with Logdata », *Proceedings of the* 19^{th} *ACM Conference on Hypertext and Hypermedia, HT 2008*, New York, NY, USA, p. 157-166, June, 2008b.

Lerman K., Plangprasopchok A., Wong C., « Personalizing Image Search Results on Flickr », *CoRR*, 2007.

Mei Q., Jiangy J., Suz H., Zhai C., « Search and Tagging : Two Sides of the Same Coin ? », *Technical Report No. 2919, University of Illinois at Urbana-Champaign (UIUCDCS-R-2007-2919)*, 2007.

Mika P., « Ontologies are us : A unified model of social networks and semantics », *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 5, n° 1, p. 5-15, March, 2007.

N.Tsimboukakis, G.Tambouratzis, « Document classification system based on HMM word map », *Proceedings of the* 5^{th} *International Conference on Soft Computing as Transdisciplinary Science and Technology*, ACM, New York, USA, p. 7-12, 2008.

Pan J., Taylor S., Thomas E., « Reducing Ambiguity in Tagging Systems with Folksonomy Search Expansion », *Proceedings of the* 6^{th} *Annual European Semantic Web Conference, ESWC 2009*, p. 669-683, June, 2009.

Rabiner L., « A tutorial on hidden Markov models and selected applications inspeech recognition », *Proceedings of the IEEE*, vol. 77(2), New Orleans, Louisiana, United States, p. 257-286, February, 1989.

S.Xu, Bao S., Fei B., Su Z., Yu Y., « Exploring folksonomy for personalized search », *Proceedings of the* 31^{st} *Annual International Conference on Research and Development in Information Retrieval, SIGIR 2008*, ACM, p. 155-162, 2008.

Trabelsi C., Jelassi N., Yahia S. B., « Scalable Mining of Frequent Tri-concepts from Folksonomies », *Proceedings of the* 16^{th} *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2012*, vol. 7302 of *LNCS*, Springer, Kuala Lumpur, Malaysia, p. 231-242, May, 2012.

Trabelsi C., Jrad A., Yahia S. B., « Bridging folksonomies and domain ontologies : Getting out non-taxonomic relations », *Procedings of the* 10^{th} *IEEE International Conference on Data Mining Workshops, ICDMW 2010*, IEEE Computer Society, Sydney, Australia, p. 369-379, December, 2010.