First International Congress on Technology, Communication and Knowledge (ICTCK 2014)
November, 26-27, 2014 - Mashhad Branch, Islamic Azad University, Mashhad, Iran

Discovery Of The Triadic Frequent Closed Patterns Based On Hidden Markov Model In Folksonomy

Maryam Fahimi
Department of Computer Engineering
Mashhad Branch, Islamic Azad University
Mashhad, Iran
Fahimi.Maryam@gmail.com

Majid Vafaei Jahan
Department of Computer Engineering
Mashhad Branch, Islamic Azad University
Mashhad, Iran
VafaeiJahan@mshdiau.ac.ir

Masood Niazi Torshiz
Department of Computer Engineering
Mashhad Branch, Islamic Azad University
Mashhad, Iran
Niazi@mshdiau.ac.ir

Abstract—**With rise of web 2.0, its associated user-centric applications have attracted a lot of users. Folksonomy plays an important role in these systems, which is made of labeling data. Discovery triadic frequent closed patterns is an important tool in knowledge discovery in folksonomy. The huge volume of data and the number of dimensions in these systems, including users, tags and resources are challenging for data mining. In this paper, a method for discovering all triadic frequent closed patterns based on Hidden Markov Model in folksonomy is proposed. By extracting useful data from dataset, the proposed method emprises to build Hidden Markov Model on the two dimensions, then with inference from created hidden model discover triadic frequent closed patternsthrough applying third dimension on the results. In fact, extracting useful data in the first step and using viterbi based algorithm, for inference, regularly are pruned dataset and are causes for triadic frequent closed patterns to be discovered more quickly. Testing on a real data set taken from "Del.icio.us" website and comparing the results with the same algorithm in the field of folksonomy called "Trias" show that the proposed method in terms of the time, can extract all triadic frequent closed patternsmore effectively.**

I. INTRODUCTION

Social resource sharing systems are web-based systems that allow users to upload their resources, and by labeling them with arbitrary words present these shared resources to others. The shared photo gallery Flickr[1], the bookmarking system del.icio.us[2]and bibsonomy[3] are examples of social tagging systems, that within a few years have acquired large number of users. The reason for success of these systems is its simple structure and ease of use that allow users to freely allocate keywords as "tag" to the resources. The result of these collaborative tagging activities in these systems leads to the user-generated classification called "Folksonomy".[1]

A folksonomy has three main parts including users, tags, and resources (triple u,t,r), in which user u assignstag t to the resource r. In these systems, the tags reflect the views of the different users about unique resourcessuch as blogs, news, photos, etc. So folksonomy contains valuable information and knowledge. The hidden knowledge discovery from folksonomy is becoming the main research task among the social sharing resources systems.

Finding shared concepts between users is one of the most important tools for extracting knowledge from folksonomy which has wide applications. For example, in [2,3,4], these concepts are used for clustering folksonomy in a web-based recommended system, and in [5,6] are used for ontology learning from folksonomy, to enhance information retrieval metrics. A triadic frequent closed pattern is a triple set (A,B,C) where each user in A has tagged each resource in C with all tags from B, and that none of these sets can be extended without shrinking one of the other two dimensions.

Our algorithm will return aset of triples, whereeach triple (A,B,C) consists of a setAof users, a setBof tags,and a setCof resources and each user in Ahas taggedeach resource inCwith all tags fromB. These triples called "tri-Concepts", and have the property that none of thesesets can be extended without shrinking one of the other twodimensions.

In this paper, a method for mining all frequent tri-Concepts fromfolksonomy is provided and shows that the proposed method discovers the patterns faster than the Trias algorithm.

The remainder of the paper is organized asfollows. Section II reviews some related works. Section IIIrecalls the key notionsused throughout this paper.In section IV, we introduce a new method for mining all frequent tri-Concepts.The empirical evidences about the performance of our approach are provided in section V. Finally, Section VI contains conclusions and future works.

[1] www.Flickr.com

[2] www.Delicious.com

[3] www.Bibsonomy.org

978-1-4799-8021-5/14/$31.00 ©2014 IEEE

II. Related Work

The discovery of shared conceptualizations opens a new research field which may prove interesting also outside the folksonomy domain: 'Closed itemset mining'. This line of research did not receive a broad attention up to now. With the rise of folksonomies as core data structure of social resource sharing systems, which formally is denoted as triple, the interest in triadic concept analysis increased again. Mining all frequent tri-concepts is thethree-dimensional version of mining all frequent closed itemsets. Indeed, folksonomy systems provide a rich resource for data analysis, information retrieval, and knowledge discovery applications. The discovery of frequent patterns is one of the important tools for knowledge extraction in these systems.

Frequent pattern mining algorithms [7,8,9] typically generate a large number of patterns and many of them are redundant. To reduce the number of frequent patterns, frequent closed pattern mining algorithms have been proposed. In [10], the A-close algorithm is proposed that uses a breadth-first search to find frequent closed patterns. CLOSET [11] and CLOSET+ [12] adopt a depth-first, feature enumeration strategy. CLOSET uses a frequent pattern tree for a compressed representation of the dataset. CLOSET+, an enhanced version of CLOSET, uses a hybrid tree-projection method to build conditional projected table in two different ways according to the density of the dataset. Both MAFIA [13] and CHARM [14] use a vertical representation of the datasets. MAFIA adopts a compressed vertical bitmap structure while CHARM enumerates closed itemsets using a dual itemset-tidset search tree and adopts the Diffset technique to reduce the size of the intermediate tidsets. Since these methods adopt a feature enumeration strategy, they cannot efficiently handle datasets with a large number of features.

Although the above-mentioned algorithms perform well in their respective application domains in 2D datasets, they cannot mine frequent closed patterns in 3D context.

The problem of finding triadic frequent closed patterns can be studied in two areas. Outside the folksonomy areas in [15], two algorithms have been proposed, the first algorithm offers a framework for usage of existing methods to discover dyadic frequent closed patterns in a three dimensional dataset. The second algorithm is called CubeMiner, and acts by recursively count of ternary relations and divides datasets into smaller parts. The CUBEMINER algorithm operates in a depth-first manner, which has the risk of causing infinite trees. Moreover,at each level, several checks are performed on each candidate to ensure its closeness and its uniqueness which is computationally very expensive.Existing several checks to achieve the unique patterns and mismatch with folksonomy is a problem in these algorithms. In [16], J¨aschke et al. introduced the TRIAS algorithm to compute frequent tri-concepts from a folksonomy. Actually, the main feature of TRIAS is to exploit the subsets of tri-concepts already extracted in order to check whether they lead to new tri-concepts. However, several tri-concepts are computed redundantly inducing a number of unnecessary computations. This drawback occurs because of the particular order of extraction of tri-concepts which is strongly inspired by the way of doing the NEXTCLOSURE algorithm. More recently, Cerf et al., in [17], proposed the DATA-PEELER algorithm with the challenge of outperforming both TRIAS and CUBEMINER algorithms in terms of performance. The DATA-PEELER algorithm is able to extract closed concepts from n-ary relations by enumerating all the n-dimensional closed patterns in a depth first manner using a binary tree enumeration strategy. In [18], Trabelsi et al. present TRI-CONS methods that directly extract triples from the folksonomy. The main thrust of the introduced algorithm stands in the application of an appropriate closure operator that splits the search space into equivalence classes for the localization of tri-minimal generators. These tri-minimal generators make the computation of the tri-concepts less arduous than do the pioneering approches of the literature.

III. PRELIMINARIES

A. Folksonomy

Folksonomies are the core structure of social bookmarking systems. The word "folksonomy" is a blend of the words "taxonomy" and "folk", and stands for conceptual structures created by the people. The way a folksonomy is emerging is the same in all these systems and can be described as follows: There is a user who is interested in a certain resource. A folksonomy system provides a way to store this resource and to annotate it [1]. This resource can be a photo, link or anything else. Typically, the annotation process is as simple as possible and is driven by keywords called tags. The set of all allocations of a user for a resource is called Personomy. Theentire personomy collections of all users is called Folksonomy. Here, a formal and precise definition of Folksonomy can help understand the issue.

Definition (1): a Folksonomy is simply a tuple F:= (U,T,R,Y):

- U, T, and Rare finite sets, whose elements are called users, tags, and resources, resp.
- Yis a ternary relation between them, i.e. $Y \subseteq U \times T \times R$ [1].

B. discovery of triadic frequent closed patterns

The frequent patterns are set of items that are repeated with a greater frequency than or equal to a threshold specified by the user. To recover the more important shared concepts we can additionally impose minimum support constraints on each of the three dimensions 'users', 'tags', and 'resources'. Definition (2) describes triadic frequent closed patterns issue.

Definition (2): Suppose that F:= (U, T, R, Y) as defined in (1) is given Folksonomy. A tri-set of F is a triple (A,B,C) with A⊆U, B⊆T ,C⊆R. also, we consider u-minsupp and t-minsupp and r-minsupp the thresholds A, B, C in such a way that u-minsupp, t-minsupp, r-minsupp∈[0,1]. The task of mining all frequent tri-concepts consists in determining all tri-sets (A,B,C) of F with$\frac{|A|}{|U|} \geq u - minsupp$, $\frac{|B|}{|T|} \geq t - minsupp$, $\frac{|C|}{|R|} \geq r - minsupp$.

Closed in a triplex pattern means that none of these sets can be extended without shrinking one of the other two dimensions.Also, sometimes it is more convenient to use absolute ratherthan relative thresholds. For this case we let τ_u:= |U|·u-minsupp,τ_t :=|T|·t-minsupp, andτ_r:= |R|·r-minsupp.[16]

IV. THE PROPOSED METHOD

In this paper, process of discovery triadic frequent closed patterns is divided into four steps and each one has been described as follows.

A. *Preprocessing*

In folksonomy, there are no boundaries to assign a tag to an item. This provides adaptability, flexibility and in the meanwhile, causes problems in analysis [3]. As a result, many rows of folksonomy dataset, may contain data that impose additional processing. Thus, in order to get useful data, we filter dataset in preprocessing step.

In the proposed method the preprocessing step consists of two phases: in the first phase only those rows of dataset where in each set of U,T,R the number of occurrences of each element is equal to multiply of other sets thresholdswill be extracted. This means that, this step will extract the data that for users as much as $\tau_r \times \tau_t$or more, to resources as much as $\tau_u \times \tau_t$ or more and for tags as much as $\tau_u \times \tau_r$or more are repeated. In the second phase, from output of previous phase, we extract a part of datasetwhere for each element in the first set, its Corresponding element in the second set is repeated as much as the threshold that is considered for third set. e.g. for each t, its related u is repeated as much as τ_r.

B. *Learning Hidden Markov Model*

A real folksonomy dataset in addition to user number, information resource address and dedicated tags may include time of tag registration, source format, tag weight, and other additional information. In this article, only user numbers, information resource address, and assigned tags, among others, are important and other information will be ignored. We only use the resources and tags data to produce HMM[4], in order to consider resource as states of HMM and tags as observations of HMM. HMM is a powerful mathematical model that obeys from the concepts of Markov models. In this section, we describe the learning HMM problem .

As mentioned in the previous sections, in the problem of miningall frequent tri-concepts, there is three set as <U,T,R> where any user in U, allocated all tags in T to each resource in R. In this problem, Q is a set of tags and is considered as observations of HMM, $Q = \{q_1, \dots, q_{nq}\}$. Also, S is a set of resources and is cosidered as states of HMM, $S = \{s_1, \dots, s_{ns}\}$.$nu_{q,j}$is the number of users who have devoted tag q to the resource j, and $\mathrm{ns}_{\mathrm{i,j}}$ is the number of times that resource j is tagged after resource i. Our HMM is as follow:

$$\lambda=(A,B,\pi) \tag{1}$$

Equation (1) is a probabilistic model which its parameters using folksonomy are computed as follows:

- $\pi = [..\pi_i..] = \mathrm{P}(\mathrm{s_i}) = \frac{\mathrm{ns_i}}{|\mathrm{Y}|}$: In this problem, the probability of be in each state as the initial factor is equal to result of dividing the number of times a tag is assigned to resource i, divided by number of rows in dataset.
- $B = [\dots b_j(q) \dots] = P(q|\mathrm{s_j}) = \frac{nu_{\mathrm{q,j}}}{\mathrm{ns_j}}$: The probability of observing tag q in state $\mathrm{s_j}$is equal to the number of times that users assigend tag q to the resource j, divided by the total times resource j is tagged.
- $\mathrm{A} = [\dots \mathrm{a_{ij}} \dots] = \mathrm{P}(\mathrm{s_j}|\mathrm{s_i}) = \frac{\mathrm{ns_{i,j}}}{\mathrm{ns_j}}$: The probability of transition from state $\mathrm{s_i}$ to $\mathrm{s_j}$is equal to the number of times that resource j is tagged after resource i, divided by the total times resource j is tagged.

Figure 1 shows an example of the proposed method HMM.

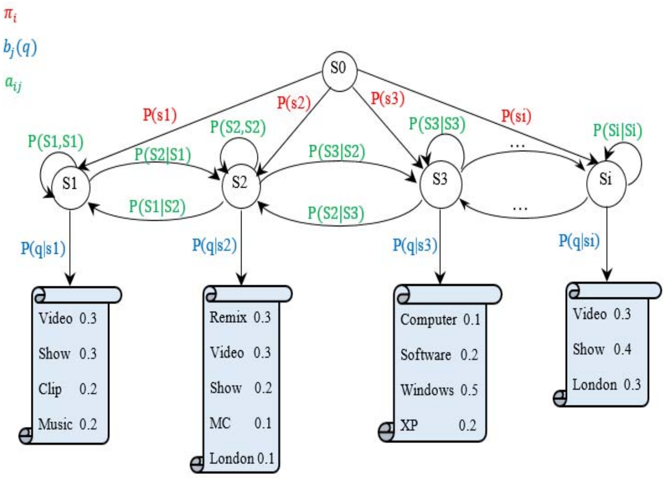

Figure 1. example of the proposed method HMM

C. *Inference of Hidden Markov Model*

To infer from created hidden model, we use aviterbi algorithm. This function receives observations sequence and HMM, then returns the probability that the observations sequence is observed in each state of HMM.

In order to support the τ_t, and mining groups of the same resources that have received similar tags, we make non repeated combinations of set t with length τ_t and consider it as observations sequence for viterbi algorithm. Then among generated outputs, only those results for later processing will be saved that at least for τ_r resources the probability of observing input string be existed. The output of this step are shared tags and resources groups that satisfy the threshold values.

D. *Final processing*

In fact , this step includes applying third dimension means that users extracted on each cluster from previous step, Which may be breaks down or join groups according to the corresponding users.

[4] Hidden Markov Model

Breaking clusters: In this process, based on each extracted dyadic frequent pattern in inference of HMM step, the clusters are broken into several clusters so that each cluster includestriple users, resources and tags and each userassigns all existing tags to each resource.

Joining clusters: After applying the third dimension and breaking clusters, clusters containing same users and resourceshaving different sets of tags may be created. This means a group of similar users are assigned two sets of tags to the same resource. In this case, in order to mine all frequent tri-conceps, we integrate two clusters on tags.

Finally,this section provides all tri-conceps.

V. Experiments

In this section, experiments and results on a real dataset are presented. The proposed method was tested on a dataset taken from Delicious website. The number of relations in this data set is |Y|=616819, which is related to mid December 2003 until June 2004. The resulting folksonomyconsists of|U| =3,301 users, |T| =30,416 different tags, |R| =22,036 resources (URLs), which are linked by |Y|= 616,819 triples. This dataset is easily downloadable5. For mining frequent tri-concepts we used minimum support values of $\tau_u := \tau_t := \tau_r := 2$ and measured the run-time of the implementations on a dual-core Opteron system with 2 GHz and 6 GB RAM.

Performance of the proposed method was evaluated on mentioned dataset and the results have been compared with the Trias algorithm. We used monthly snapshots as follows: F_0contains alltag assignments performed on or before Dec 15, 2003, together with the involved users, tags, and resources; F_1all tag assignments performed on or before Jan 15, 2004, together with the involved users, tags, and resources; and so on until F_6which contains all tag assignments performed on or before June 15, 2004, together with the involved tags, users, and resources, Which is shown in detail in Table 1.

TABLE I. Detail of datasets

DataSet	**|R|**	**|T|**	**|U|**	**|Y|**
F0	54043	5670	588	98870
F1	65994	8290	941	131968
F2	89709	12599	1472	201935
F3	115167	16357	1843	278277
F4	146197	20970	2292	380804
F5	182125	25424	2795	493380

As shown in Table 2, the number of extracted patterns in the proposed method is exactly equal to the Trias, but, the proposed method in terms of the time is faster and can more effectively extract triadic frequent closed patterns.

TABLE II. Performances of proposed method vs. Trias

DataSet	**Proposed Method**		**Trias**	
	Number of patterns	***Run time (in seconds)***	***Number of patterns***	***Run time (in seconds)***
F0	25	0.8	25	5
F1	53	1.8	53	10
F2	129	8	129	46
F3	266	36	266	116
F4	491	113	491	315
F5	734	246	734	691
F6	1141	684	1141	1410

Note that through removingunnecessary data in the preprocessing step time consuming processes have been avoided. This step dramatically increases the speed in creation and inferences from the HMM. In fact, whatever the input threshold values in preprocessing step chosen larger, it identifies more non-useful data and reduces the size of the input data to creation and inference of HMM, and are causes for runtime to be much lower, and conversely. In folksonomy due to the large number of users, the extent of resources and lack of restrictions on the choice of the tag, the more relations are isolated. For example, by considering $\tau_u := \tau_t := \tau_r := 1$ after discovery all frequent tri-concepts more sets of three parts (A, B, C) are in fact the same relations in folksonomy. But the important thing is that choosing small values as mentioned above is not consistent with the logic of discovering frequent closed patterns, since the goal is to find overlapping sets. This can be achieved by choosing larger values for the parameters. However, what can be stated as a characteristic of the proposed method is that the performance of the proposed method is influenced by the choice of parameters. Figure 2 shows the effect of chosen various parameters.

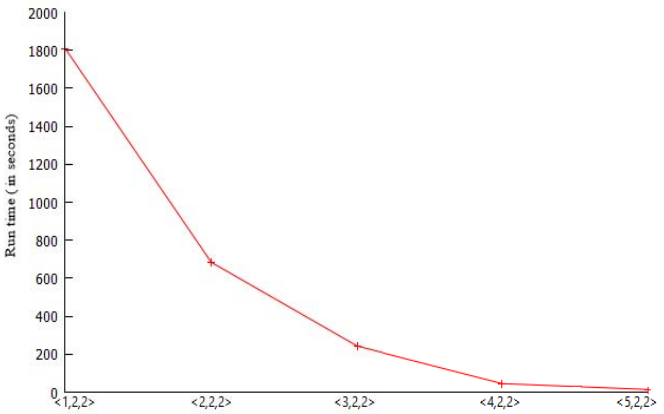

Figure 2. Threshold parameters<τ_u,τ_t,τ_r>

In this test, threshold of tags and resources are fixed and the threshold of usersin each run is increased by one unit. It is

[5] http://www.uni-koblenz-landau.de/koblenz/fb4/AGStaab/Research/DataSets/PINTSExperimentsDataSets/index_html

clearthatby increasing the threshold for the user, execution time is reduced.

Conclusions and Future Work

In this paper, a method for discovering triadic frequent closed patterns based on Hidden Markov Model in folksonomy is presented. Existing methods are incapable of discovering triadic frequent closed patterns or are not closely matchedwith the scope and range of folksonomy. Experimental results show that the proposed method is fully capable of solvingthis problem and is quicker than similar algorithm in the field of folksonomy called trias. Since the main problem in discovering frequent closed patterns matter is time, in future work, our main focus is on minimizing runtime in solving this problem. This method can also be used on datasets with more dimensions, such as when the user location or time is intended. This makes it possible to obtain valuable information on folksonomy systems. Moreover, the extracted patterns can be well used in the recommended systems.

References

[1] R. J¨aschke, A. Hotho, C. Schmitz, B. Ganter, G. Stumme, " Discovering shared conceptualizations in folksonomies ", Web Semantics: Science, Services and Agents on the World Wide Web 6, pp. 38-53, 2008.

[2] I. Cantador, I. Konstas, J. Jose, "Categorising social tags to improve folksonomy-based recommendations", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 9, pp. 1-15, 2011.

[3] Y. Ustunbas, S. G. Oguducu, "A Recommendation Model for Social Resource Sharing Systems Based on Tripartite Graph Clustering", Intelligence and Security Informatics Conference (EISIC), European, pp. 378-381, 2011.

[4] T. Yoshida, U. Inoue, "A Bookmark Recommender System Based on Social Bookmarking Services and Wikipedia Categories",ACIS 14th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 409-413, 2013.

[5] C. Wen-hao, C. Yi, L. Ho-fung, L. Qing, "Generating ontologies with basic level concepts from folksonomies", International Conference on Computational Science, ICCS, vol. 1, pp. 573-581. 2010.

[6] C. Trabelsi, A. Ben Jrad, S. Ben Yahia, "Bridging Folksonomies and Domain Ontologies: Getting Out Non-taxonomic Relations", 2010 IEEE International Conference on Data Mining Workshops, pp. 369-379, 2010.

[7] R.Agraval, R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", VLDB '94 Proceedings of the 20th International Conference on Very Large , Morgan Kaufmann Publishers Inc, pp. 487-499, 1994.

[8] P. Shenoy, J. R. Haritsa, S. Sudarshan, G. Bhalotia, M. Bawa, D. Shah, "Turbo-charging vertical mining of large databases", SIGMOD Rec, vol. 29, pp. 22-33, 2000.

[9] M. J. Zaki, S. Parthasarathy, M. Ogihara, W. Li, "New Algorithms for Fast Discovery of Association Rules", University of Rochester, 1997.

[10] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, "Discovering Frequent Closed Itemsets for Association Rules", ICDT '99 Proceedings of the 7th International Conference on Database Theory, pp. 398-416, 1999.

[11] J. Pei, J. Han, R. Mao, "Closet: An efficient algorithm for mining frequent closed itemsets", (ICDT□99)Proceeding of the 7th international conferenceon database theory, Israel, pp 398-426, 2000.

[12] J. Wang, J. Han, J. Pei, "CLOSET+: searching for the best strategies for mining frequent closed itemsets",ACM SIGKDD Proceedingsof the ninth international conference on Knowledge discovery and data mining, pp. 236-245, 2003

[13] D. Burdick, M.Calimlim, J. Gehrke, "MAFIA: a maximal frequent itemset algorithm for transactional databases", 17th International Conference on Data Engineering, pp. 443-452, 2001.

[14] M. Zaki, C. Hsiao, "CHARM: An efficient algorithm for closed association rule mining", In SDM'02, Arlington, VA, USA, pp. 457-473 April 2002.

[15] L. Ji, K. Tan, A. K. H. Tung, "Mining frequent closed cubes in 3D datasets", VLDB '06 Proceedings of the 32nd international conference on Very large data bases, Seoul, Korea, pp. 811-822, 2006.

[16] R. Jaschke, A. Hotho, C. Schmitz, B. Ganter, G. Stumme, "TRIAS--An Algorithm for Mining Iceberg Tri-Lattices", Proceedings of the Sixth International Conference on Data Mining (ICDM'06), pp. 907-911, 2006.

[17] L. Cerf,J. Besson, C. Robardet, J. Boulicaut, "Closed patterns meet n-ary relations", ACM Trans. Knowl. Discov. Data, vol 3, pp. 1-36, 2009.

[18] C. Trabelsi, N. Jelassi, S. Ben Yahia, "Scalable Mining of Frequent Tri-concepts from Folksonomies", Springer Berlin Heidelberg, vol 7303, pp. 231-242, 2012.

[19] A. Hotho, "Data Mining on Folksonomies", Springer Berlin Heidelberg, vol. 301, pp. 57-82, 2010.