# Folksodriven Structure Network

Massimiliano Dal Mas

*me @ maxdalmas.com*

**Abstract.** Nowadays folksonomy is used as a system derived from user-generated electronic tags or keywords that annotate and describe online content. But it is not a classification system as an ontology. To consider it as a classification system it would be necessary to share a representation of contexts by all the users. This paper is proposing the use of folksonomies and network theory to devise a new concept: a "Folksodriven Structure Network" to represent folksonomies. This paper proposed and analyzed the network structure of Folksodriven tags thought as folsksonomy tags suggestions for the user on a dataset built on chosen websites. It is observed that the Folksodriven Network has relative low path lengths checking it with classic networking measures (clustering coefficient). Experiment result shows it can facilitate serendipitous discovery of content among users. Neat examples and clear formulas can show how a "Folksodriven Structure Network" can be used to tackle ontology mapping challenges.

**Keywords**. Ontology, Folksonomy, Natural Language Processing, Information filtering, Metadata, Semantic networks, Scale-free Network, Reasoning, Algorithms, Experimentation, Theory

## 1 Introduction

The communicative form of the World Wide Web is based on user-centric publishing and knowledge management platforms as sharing systems for social and collaborative matching like: Wikis, Blogs, Facebook, etc… Ontology defines a common set of sharing concepts [1], but unfortunately ontologies are not wide spread at the moment. While Folksonomy is said to provide a democratic tagging system that reflects the opinions of the general public, but it is not a classification system and it is difficult to make sense of [2]. A representation of contexts should be share by all the users. The goal of this work is to help the users to choose proper tags thanks to a "Folksodriven Structure Network" intended as a dynamical driven system of folksonomy that could evolve during the time. In this work the main network characteristics was analyzed by a group of articles from a chosen websites and analyzed according to the Natural Language Processing. The data structures extracted is represented on folksonomy tags that are correlated with the source and the relative time exposition - measure of the time of its disposal. Considering those we define a tag structure called Folksodriven, and adapt classical network measures to them.
An extension of this paper is available on Arxiv ( http://arxiv.org/abs/1109.3138 ).

## 2 Folksodriven Notation

$$FD := (C, E, R, X) \quad (1)$$

A Folksodriven will be considered as a tuple (1) defined by finite sets composed by:

- *Formal Context* (C) is a triple *C:=(T, D, I)* where the objects *T* and the attributes *D* are sets of data and *I* is a relation between *T* and *D* [3] – see 3 (*Folksodriven Data Set*);
- *Time Exposition (E)* is the clickthrough rate (CTR) as the number of clicks on a *Resource (R)* divided by the number of times that the *Resource (R)* is displayed (impressions);
- *Resource (R)* is represented by the uri of the webpage that the user wants to correlate to a chosen tag;
- *X* is defined by the relation $X = C \times E \times R$ in a Minkowski vector space [4] delimited by the vectors *C*, *E* and *R*.

## 3 Folksodriven Data Set

The data set has been built from articles taken from web sites news for a period of one month, because they are frequently updated. Tokens will be extracted from the title (*T*) and the description (*D*) of the articles. Those tokens compose a data set of words proposed to the users as Tags that he/she can add to a document - the articles on the web sites - to describe it. Chunking was used in this work as a starting point but it is at a very low semantic level.
In this paper the notion *context* is used in the sense of *formal context* as used in the ontological sense defined by the Formal Concept Analysis (FCA) - a branch of Applied Mathematics [5] - for the dynamic corpus on chunking operation. A *set of formal contexts C* is defined by (2) considering: *T* as a *set of title tags, D* as a *set of description tags, I* as a *set of incidence relations of context* defined by the frequency of occurrence of the relation between *T* and *D* as depicted in (3). The tag *T* derived by the title was considered as a facet described by the tag *D* derived by the description. On (2) the *set of incidence relations of context I* is defined by the matching between *T* and *D* tags by relation (3) allowing multiple associations among *D* tags and the faceted context defined by every *T* tag.
Multiple matching was disambiguated by updating a Jaccard similarity coefficient [6] associated with the incidence relation of context. In this way a selected number of chunks, defined according to the *Formal Context* (*C*), are proposed to the user as folksonomy tags for the correlated uri *Resource* (*R*). So the "Folksodriven Data Set" can "drive" the user on the choice of a correct folksonomy tag.

$$C_n := (T_n, D_n, I_n) \quad (2) \qquad I \subseteq T \times D \quad (3) \qquad K_r(i) := \frac{|C_r(i) \times E_r(i)|}{|C_r(i)| \bullet |E_r(i)|} \quad (4)$$

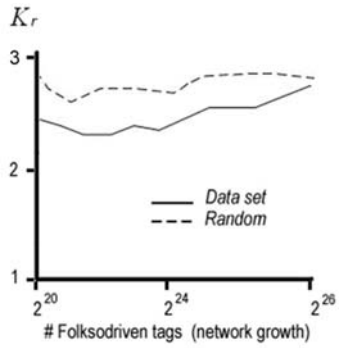


**Figure 1:** Set of data compared with the corresponding Random graphs for Folksodriven *Clustering Coefficient*. It is depicted how the characteristic path length takes quite similar values for the corresponding Random graph.

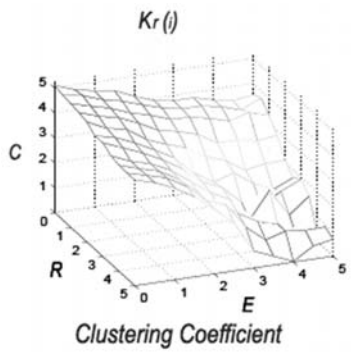


**Figure 2:** *Clustering coefficient* is depicted in the space delimited by *C*, *E*, and *R*. For larger time exposition *E* the *Clustering Coefficients* become drastically smaller, as expected for the $E \rightarrow \infty$ and $C \rightarrow 0$ limit.

## 4 Folksodriven as a Network

The Folksodriven tags can be depicted by network patterns in which nodes are Folksodriven tags and links are semantic acquaintance relationships between them according to the SUMO (http://www.ontologyportal.org) formal ontology that has been mapped to the WordNet lexicon (http://wordnet.princeton.edu). It is easy to see that Folksodriven tags tend to form groups (as small groups in which tags are close related to each one) and Folksodriven tags of a group also have a few acquaintance relationships to Folksodriven tags outside that group. Folksodriven tags may be considered the hubs responsible for making such network a "Scale-free Network". In a Scale-free Network most nodes of a graph are not neighbors of one another but can be reached from every other by a small number of hops or steps, considering the mutual acquaintance of Folksodriven tags [7, 8].
An important characteristic of Scale-free Networks is the *Clustering Coefficient* distribution, which decreases as the node degree increases following a power law [8]. We consider an exclusive vs. an overlapping clustering as the ratio between the maximum and the minimum value connectedness of the neighbours of a Folksodriven tag to the uri resource *r* considered (4).

## 5 Experiments

A test network model was realized in a simulated environment to check the Scale-free Network structure of the Folksodriven tags. The Scale-free Network was compared with a random graph generated adding tags one at a time joining to a fixed number of starting tags, that are chosen with probability proportional to the graph degree - model developed by Barabasi and Albert [8]. All data were obtained from averages over 100 identical network realizations with a sample of 400 nodes taken randomly from each graph performing twenty runs to ensure consistency. The *Clustering Coefficient* has remained almost constant at about 2.5 while the number of nodes has grown about twenty during the observation period. On average, every *Formal Context* (C), *Time Exposition (T)* and *Resource (R)* defined on the original data set can be reached within 2.5 mouse clicks from any given page. This attest the context of "serendipitous discovery" of contents in the folksonomy community [9].

**Massimiliano Dal Mas** is an engineer at the Web Services division of the Telecom Italia Group, Italy. His interests include: user interfaces and visualization for information retrieval, automated Web interface evaluation and text analysis, empirical computational linguistics, and text data mining. He received BA, MS degrees in Computer Science Engineering from the Politecnico di Milano, Italy. He won the thirteenth edition 2008 of the CEI Award for the best degree thesis with a dissertation on "Semantic technologies for industrial purposes" (Supervisor Prof. M. Colombetti).

## References

[1] M. Dal Mas (2010). *Ontology Temporal Evolution for Multi-Entity Bayesian Networks under Exogenous and Endogenous Semantic,* CORR - Arxiv ( http://arxiv.org/abs/1009.2084 )
[2] E. K. Jacob (2004). *Classification and categorization: a difference that makes a difference*
[3] G. Stumme (1999). *Acquiring expert knowledge for the design of conceptual information systems*, in Proc. 11 th. European Workshop on Knowledge Acquisition", Dagstuhl Castle, 1999, pp. 275-290
[4] F. Catoni, D. Boccaletti, R. Cannata (2008). *Mathematics of Minkowski Space*, Birkhäuser, Basel.
[5] K. Shen, L. Wu (2005). *Folksonomy as a Complex Network*, CORR
[6] T. Pang-Ning, M. Steinbach, V. Kumar (2005). *Introduction to Data Mining*, Boston: Addison-Wesley
[7] C. Cattuto, C. Schmitz, A. Baldassarri, V. D.P. Servedio, V. Loreto, A. Hotho, M. Grahl, G. Summe. (2007). *Network properties of folksonomies* Proceedings of the WWW2007
[8] A. Vazquez, J Oliveira, Z. Dezso, K Goh, I. Kondor, A. Barabási (2006). *Modeling bursts and heavy tails in human dynamics*. Physical Review E 73(3), 2006
[9] M. Dal Mas (2011). *Folksodriven Structure Network*, CORR - Arxiv ( http://arxiv.org/abs/1109.3138 )