Multimed Tools Appl (2016) 75:573–605
DOI 10.1007/s11042-014-2309-3

Semantics discovery in social tagging systems: A review

Fouzia Jabeen · Shah Khusro · Amna Majid · Azhar Rauf

Received: 14 February 2014 / Revised: 18 August 2014 / Accepted: 6 October 2014 /
Published online: 18 October 2014

Abstract Web 2.0 has brought many collaborative and novel applications which transformed the web as a medium and resulted in its exponential growth. Tagging systems are one of these killer applications. Tags are in free-form but represent the link between objective information and users' cognitive information. However, tags have ambiguity problem reducing precision. Hence search and retrieval pose a challenge on folksonomy systems which have flat, unstructured, non-hierarchical organization with unsupervised vocabulary. We present a brief survey of different approaches for adding semantics in folksonomies thus bringing structure and precision in search and navigation. We did comparative analysis to estimate the significance of each source of semantics. Then, we have categorized the approaches in a systematic way and summarized the feature set support. Based on the survey we end up with recommendations. Our survey and conclusion will prove to be relevant and beneficial for engineers and designers aiming to design and maintain well structured folksonomy with precise search and navigation results.

Keywords Semantics in folksonomy · Folksonomy enrichment · Search precision · Navigation

1 Introduction

This new period of the Web, also recognized as Web 2.0, has brought a diversity of new social applications like wikis, blogs, social networks, social bookmarking, photo, music and video sharing sites, bringing into existence many collaborative and novel applications which are highly accepted among users and are very successful. These applications made it possible for all users of web to add and share huge amounts of multimedia content, and to label these content resources with free-form keywords commonly called tags.

Web sites such as Flickr and YouTube, called the tagging applications, support users to tag user-generated photos and videos. In comparison, Amazon and Del.icio.us motivate users to give tags to products or existing web pages. These tagging processes led to the emergence of folksonomies where the tags are freely created by users keeping in mind the context in which the user is tagging a resource. According to some authors, the widespread use is accredited to two main factors, firstly, tags are very simple and easy to create; the users do not need any

F. Jabeen (✉) · S. Khusro · A. Majid · A. Rauf
Department of Computer Science, University of Peshawar, Peshawar 25120, Pakistan
e-mail: fouzia.jabeen@gmail.com

particular skills or experience to tag. They are easy to use for all users even if they have different level of understanding, age, cultural backgrounds and languages. It requires no setup and is very easy to adapt. Secondly, tagging is instantaneous [38]. Furthermore, folksonomies do not require any hierarchy or other classification scheme that's why it is open-ended and truly reflects user perspective regarding different resources. It involves low cognitive cost [77]. Users have freedom in assigning tags that they think are suitable for a resource and this freedom of following their own vocabulary is the basic reason behind the success of tagging systems. Users utilize tags to retrieve or explore information, to add or share resources, to catch the attention of other people, to introduce themselves in a community, or to convey their opinion [38].

Information retrieval is very important in searching databases that are tag based, as there is large number of different kinds of resources with variable number of free form tags assigned to them. Folksonomies are very useful attempt to improve precision in searching and retrieving information. When different users assign metadata to a web resource in the form of tags, users' consensus in the form of user generated classification emerges automatically [74]. Because of this consensus, users can find unexpected information that they didn't know but is relevant to them [77].

The liberty and freedom, however, leads to the problem of highly unstructured tags. Tag meanings get ambiguous due to spelling mistakes, different lexical forms of the same word (morphological variation), polysemy, homography, synonymy, detail/granularity level, multi-lingualism, inaccurate tag-to-resource associations, different levels of tag precision and abstraction [8, 44, 77, 80]. Due to these reasons, tag space is inconsistent, inefficient and noisy. This reduces precision and recall in search results. As folksonomy has a flat organization having no explicit semantic relations among tags [44, 112], it is difficult to find relevant tags and to navigate through them. Due to this unstructured form, tagging in folksonomy poses a serious challenge to information retrieval. Current systems pay no attention to resources tagged with morphological variations or synonyms of that tag, as well as the resources tagged with more generic or more specific tags, or the same tag written in another language. In addition, when searching with polysemous tags, all the resources tagged with that tag are retrieved without considering the sense of the tag, the user was looking for [38].

By making different semantic relations (like equivalence, subsumption etc.) explicit [63] and at different abstraction levels, it will be easy to locate the tags. In addition, it will also show the level of generality or specificity of tags. Furthermore, when the user enters search keywords, these may not be specifically from the domains that a folksonomy covers. There exist tags that are different in scope but are very relevant. So, they must be disambiguated independent of their domains. A system solution for folksonomy problems may be developed by disambiguating tags and arranging them in some hierarchical structure (at different granularity levels in the form of tag bundles or in some other representation). The tag space can be further enriched with different novel features like bursty events and tags, enriching it with more metadata (secondary tags in addition to the primary tags) to increase precision and recall ratios and removing spam posts for correct tag to resource association.

This paper presents an extensive survey of different approaches based on the previously mentioned aspects and other semantic emergent features. The major contributions of the paper are following:

- We did comparative analysis of semantic incorporating sources to estimate significance of each source and to highlight their strengths and limitations.
- We have categorized the recent and state-of-art semantic emergent methods that result in precision in search and navigation.

- We have summarized these methods highlighting their accuracy, pros/cons and supported feature list.

The paper is organized in seven sections according to Framework shown in Fig. 1. In section 2, we are discussing significance of semantics and semantic incorporating sources (External, Mathematical/statistical formulas, and folks) utilized by different researchers for introducing semantics in folksonomy. Making these sources and an additional Hybrid category (Combination of statistical, knowledge based and folk) as basis, recent and state-of-art semantic emergent methods for *Bringing structure*, *Protection of folksonomy structure* and *Enriching query along with search results* are categorized in Section 3, 4 and 5 respectively. We are focussing on these three aspects (*Bringing structure*, *Protection of folksonomy structure* and *Enriching query along with search results*) because they are interrelated and have the same objective of achieving precision in navigation and searching. Some of the techniques are not folksonomic but in our opinion can effectively be utilized in tagging model. These are discussed under the category '*Other Aspects*'. Lastly in section 6, we have presented the summary and ended up with conclusion and recommendations in section 7.

2 Semantics in folksonomy

There is a lot of work done in order to introduce semantics in folksonomy. Braun et al. [18] compared some novel web applications that provide semantic tagging and thus, result in increase precision. The work of Yeung et al. [124] is based on mutual contextualization of users, tags and resources. They analysed semantics emerging from the bipartite graphs for all three elements (the users, tags and resources) of folksonomy.

A survey based on social tagging techniques by M.Gupta et al. [44] discussed models of tag generation, user motivations for tagging, tag space visualization, aspects of ambiguity removal, hierarchy generation and spamming. J. Trant et al. [112] presents another review which discusses research literature on folksonomies till 2007. In this section let's have a look at the sources utilized for semantic induction and its significance.

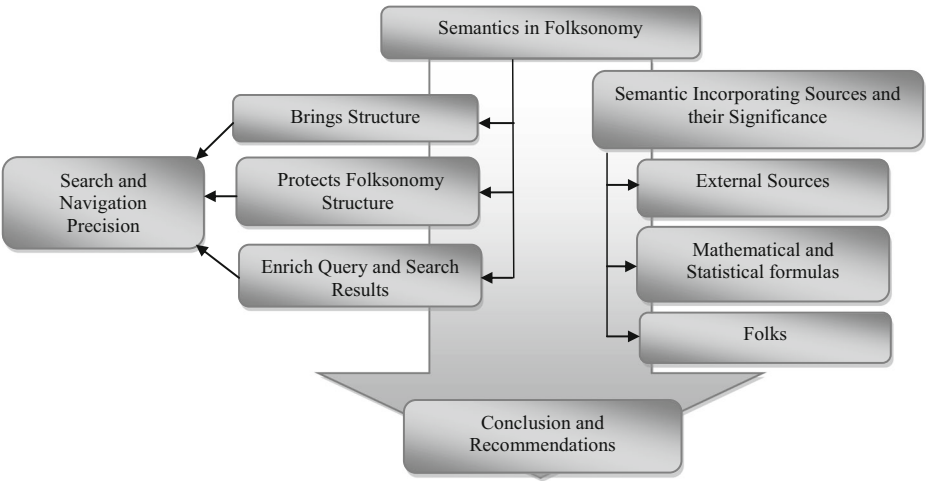


Fig. 1 Framework

2.1 Significance and role of semantic sources

Semantics in folksonomy can be incorporated utilizing different sources; the prominent ones are Knowledge based sources like Wikipedia/DBpedia, Ontologies (collectively called external sources), Statistical/mathematical formulas, and Folk perspective. Let's have a look at each source and its significance.

2.1.1 External/Knowledge sources

Wikipedia Wikipedia is one of the finest examples of collaboratively created and crowd sourcing based content on the web. According to Alexa.com[1] Wikipedia is among the top 10 sites most visited on the web. Other wiki based online encyclopaedias like Scholarpedia[2] and Citizendium[3] are also available, however, they allow registered users only.

There have always been questions regarding quality of Wikipedia due to its open to edit nature. So, many approaches have been used to prove it as a reliable data source. Wikipedia can be considered a source of information as reliable as Britannica, analysed and stated by Jim Giles [41]. In order to assess the quality of Wikipedia articles, Kittur et al. [56] used article assessment project of Wikipedia in which articles were assigned grades analysing how much real facts they contain, and how much accurate, verifiable, unbiased, stable and comprehensive they are. They validated it externally by non-Wikipedian community too. Results from external community were also highly significant. Javanmardi et al. [52] compared registered and non registered users of Wikipedia to statistically assess the quality of their contribution in editing wiki text. Results showed that most of the changes in this online encyclopaedia are made by the registered users and the ones made by the non-registered users are in a short number. Data resulting from Wikipedia articles is not biased and is validated by collective intelligence of editors worldwide.

Currently, extensive research has been done on utilizing Wikipedia or its RDF (Resource Description Framework) form called DBpedia as a data source, and taking advantage of this collaborative effort.

DBpedia DBpedia is the Semantic Web version of Wikipedia [34]. We can use it to ask sophisticated queries against Wikipedia. Bizer et al. [16] and Auer et al. [6] DBpedia project extracts structured information from Wikipedia so that semantic web techniques can be applied on it. As Wikipedia evolves, changes in Wikipedia are reflected in DBpedia. So, it is continuously updated. Thus, the problems like non machine understandability, non-freshness and topic coverage can be covered by DBpedia.

WordNet WordNet is a well-organized taxonomic knowledge base and in many researches has been utilized for finding semantic relatedness. It consists of both lexical units and the relations among them, structured into a relational semantic network. Basic intention for its development was to create a product that could merge the advantages of electronic dictionaries and on-line thesauri. Thus, making it an ideal tool for disambiguation of meaning, semantic tagging and information retrieval. In WordNet each distinct meaning of a word is presented by a *synset*. Synsets are linked to each other through explicit semantic relations (synonymy, antonymy, is-a, part-of, etc.). This creates a network where related concepts can be recognized by their relative

[1] http://www.alexa.com/
[2] http://www.scholarpedia.org/
[3] http://en.citizendium.org/

distance from each other. It is outlined by Wu and ZHOU [116] that for nouns, the most common, important and useful relation is 'is-*a* relation'. It covers over 70 % of the total relations that exist for nouns. This relation is covered in WordNet. To achieve the goal of *Multilingual WordNet,* one of the most significant attempts is *EuroWordNet,* whose ultimate aim is to develop multilingual databases with WordNets for several European languages.

Wikipedia and WordNet are created with different objectives and both have been used as powerful semantic incorporating sources. In some researches, these two sources are compared to highlight their strengths and limitations. Haridas et al. [47] state in their work that discrete knowledge bases like IMDB(Internet Movie Database) and WordNet do not cover very diverse topics. To explore topics of interest that are very new and diverse (cannot be rightly classified in existing categories), we require other knowledge sources. Strube et al. [105] verified that Wikipedia computes semantic relatedness better than WordNet and Google Counts baseline. They did experiments comparing WordNet and Wikipedia on different benchmarks by applying WordNet based measures to compute semantic relatedness on Wikipedia.

Ontology In relation to search and browsing limitations, ontology solves two major problems recognized in folksonomies: (1) tag variety, similar verb tenses, plurals, spellings, synonyms etc., and (2) the different aims or types of tags used by the users, taking into consideration a separation between personal and common tags. Information retrieval becomes rich by introducing the ontology in folksonomy as it solves the problem of ambiguity and tag explosion [32]. According to Braun et al. [17] ontologies face challenge of evolving data and work process. To achieve ontology-based sustainable systems, ontology building should be done by people having domain knowledge and not just by knowledge experts.

2.1.2 Statistical and mathematical techniques

The simple and effective approaches utilized by many researchers in bringing semantics to folksonomy are based on mathematical and statistical formulas. Mathematical and Statistical formulas play an important role. The best thing about them is they are clear and unambiguous. One important reason as observed by Cattuto et al. [20] is that the vocabulary of folksonomies contains lots of community- specific terms, which are not present in any lexical resource. Thus, value is given to the utilization of distributional measures in folksonomies as compared to mapping tags to a thesaurus.

Aschke et al. [5], observed that many factors limit WordNet from extensive coverage of Del.icio.us tags. WordNet only provides coverage of English language and is composed of static body of words while Del.icio.us has tags from different languages. In addition, tags are not considered as words at all, rather considered as string of characters in Del.icio.us. Another restrictive factor is the structure of WordNet since at maximum only 61 % of 10000 most repeated tags in Del.icio.us can be found in WordNet. These facts encourage the use of statistical and mathematical techniques.

2.1.3 Folks

Research shows that user's tagging motivation is the key factor for success of tagging systems. The web demo [3] stresses the need, besides content, to know more about the user's intent in order to improve search. Ko¨rner et al. [57] also state that collective intelligence will be more precise if tagging pragmatics can also be analysed. They say that as users are the basic factor behind evolution of semantics in folksonomy, there will be some specific composition of

crowd that contributes maximally to semantics emergence in folksonomy. They differentiate among different folks (folksonomy users) as categorizers or describers. The distinction they identified represents folk's pragmatic behaviour in the sense that how much they contribute to emerging semantics. They identified and showed experimentally a specific group of taggers that add semantic precision in folksonomy. User tagging behaviour (tagging actions) shows interest and perceptions of users for different tags [115].

In this section, we have briefly discussed the significance of semantic incorporating sources. In the next section, we will focus on the techniques using these sources for bringing structure, maintaining the structure by protecting it from spam and enriching query.

3 Utilizing sources of semantics to bring structure in folksonomies

Organization brings structure. In this section, keeping different types of semantic relationships as organizational criteria, we have categorized the semantic discovery techniques. Choudhury et al. [25] used statistical and/or external knowledge based classification. However, we have classified by adding folks and hybrid based classification as well.

3.1 Similarity/Equivalence

Researches viewed and evaluated similarity in various ways by finding similarity among tags, tag to resource(s) similarity/association, resource to resource and user to user similarity. Let's have a look at the approaches.

- *Statistical / Mathematical Approaches* - Classical metrics to find similarity between any two tags $tag1$ and $tag2$ include cosine, jaccard and dice as given in Eq 1 and Eq 2 and Eq 3 respectively. However, cosine seems to yield more synonyms and siblings [105].

$$\cos ine(tag1, tag2) = \frac{tag1 \cdot tag2}{\|tag1\| \cdot \|tag2\|} \tag{1}$$

$$jaccard(tag, tag2) = \frac{|tag1 \cap tag2|}{|tag1 \cup tag2|} \tag{2}$$

$$dice(tag1, tag2) = 2\frac{|tag1 \cup tag2|}{|tag1| \cup |tag2|} \tag{3}$$

  Markines et al. [76] focused on tag to tag and resource to resource similarity. They used different methods of aggregation (projection, distribution, incremental, collaborative filtering) and evaluated them against similarity measures like cosine, overlap, jaccard and mutual information. In non-incremental methods, distributional and mutual information performed best. Same was the case for incremental method. Furthermore, the approach is verified by using WordNet for tags similarity and Open Directory Project (ODP) for resource similarity.

  Mousselly et al. [87] proposed an approach called Adapative Jenses-Shannon Divergence (AJSD) for finding related tags and is based on calculating distance between tag distributions using Jenses-Shannon Divergence. Probability distribution for each tag

is calculated using co-occurrence and Laplacian. The authors evaluated their scheme using WordNet and compared it with cosine similarity.

Combination of morpho-syntactic and semantic similarity measures are proposed by Geir and Atle [39]. Levenshtein distance for morpho-syntactic similarity while tag signatures and cosine similarity have been used to find the semantic similarity among tags. No external linguistic resources (WordNet or even semantic resources like ontologies) have been used to mine tag pairs, making this approach more robust in terms of handling a larger portion of the tags found in the folksonomy. In addition, proposed approach does not necessarily depend on tags to co-occur for finding relations among them, rather it is focused at using topical/semantic similarity in addition to the Levenshtein distance for finding similar tags.

Quattrone et al. [93, 94] argued and emphasized that real world folksonomies are characterized by power law distributions of tags, over which commonly used similarity metrics, including the Jaccard coefficient and the cosine similarity, fail to compute. Mutual reinforcement principle has been proposed which states, "*two tags are deemed similar if they have been associated to similar resources, and vice-versa that is, two resources are deemed similar if they have been labelled by similar tags*", in order to compute tag and resource similarity in large-scale folksonomies.

SHIATSU is a system developed by Bartolini et al. [10] for automatic suggestion of user labels for videos at the shot level. SHIATSU is based on the opinion that the objects that share similar visual content also have the same semantic content. This leads to conclusion that content wise similar objects should be tagged using the same set of labels. One important aspect that can influence selection of candidate set of tags (to be assigned to a resource) based on considering tags of content wise similar resources is tagging behaviour. Golbeck et al. [43] worked on examining the tagging behaviour with respect to image content. One of the important conclusions they highlighted is that, the users give more tags to images that are more visually complex. However, number of tags decrease when the numbers of Areas of interest (AOIs) exceed a certain threshold.

For recommendation purpose Lops et al. [71] computed set of candidate tags using content and collaborative components. Collaborative part is based on the analysis of tags assigned to most similar resources (about same topic), while the content-based part exploits the content of the resources that is the information emerging from contents of the resource (Content based Tagging). Based on the same idea Zhou et al. [133] presented their hybrid probabilistic model (HPM) which combines low level image features and user provided tags (Content based and Collaborative tagging) to provide appropriate tags to label images.

Distributional Measures discover the similarity among tags keeping in consideration the resource, tag and folk [21, 51]. In *Resource Context Based approach* the context of a tag tag_i considers all the resources that are annotated with tag tag_i. Abbasi [1] formally, represents the resource context of a tag tag_i as a resource vector R as shown in Eq 4.

$$R = [f\_\{ij\}] \tag{4}$$

Where f represents number of times tag tag_i appeared with the resource j. Each row of matrix R represents tag vector and each column of matrix R stands for a resource vector. Non-zero elements give count of number of times the resource has been annotated with a particular tag and zero value represents tags not used. To find tags tag_i and tag_j that are semantically similar based on their resource context, first, compute resource context R for

each of tag tag_i and tag_j. Then cosine, dice, jaccard, probabilistic (Mutual Information) and heuristic can be used to compute the similarity between resource vectors.

In Folk Context Based approach the user context of a tag consists of all the users that share identical tags. For example, if many users annotate different resources with the tags coin and cent and they do not use these two tags together in any of the resources they annotate, it would still be likely to discover relationships that exist in these tags by taking into consideration all the users that have both of these tags in common. The user context of a tag tag_i as a vector u is computed as given in Eq 5 [1].

$$U = [u\_\{ij\}] \tag{5}$$

If a user j has utilized the tag tag_i, value of u will be 1, otherwise u will be 0. Each row of matrix U is a tag vector while each column of the matrix U is a user vector. Non-zero values stand for the users that have used particular tag. Similarity between two tag vectors based on the user context can be computed using cosine, dice, jaccard, probabilistic (Mutual Information) and Heuristic.

In *Tag Context Based approach* two tags are considered similar if they occur in the same context.Tag context similarity is scalable and accurate tag similarity measurement as pointed in [21, 76].Tag context similarity is utilized by Benz et al. [13] by taking Flicker and Del.icio.us folksonomies to measure tag similarity at a global scale. As many of the frequently occurring Del.icio.us tags also appear in Flickr. The assessment of tags across Flickr and Del.icio.us shows little semantic overlapping, being tags in Flickr related more to visual point of view whereas in Del.icio.us they are inclined more towards their technical meaning. Tags can be contextualised in a better way by taking into account the social contexts in which they appear, believed by Yeung et al. [125]. While a tag itself offers slight information on this, its associations with other tags, users and documents in a folksonomy provide valuable clues for understanding its semantics.

The Tripartite Topic Model (TTM) model is applied on folksonomy by Harvey et al. [48], to put forward new tags to users (keeping in view a small number of tags that they have given) as well as their previous annotations. This model suggests more appropriate tags than current systems. TTM provides a complete representation of the data acquired from a folksonomy and so could be applied effortlessly on useful estimations such as to find similar user groups by clustering. The tag recommendation algorithm could be tailored to propose new resources instead of tags. Xu et al.[119] state an important point that in reality most of the tags are inappropriate to image content. Solutions presented in many researches are based on tag similarity in order to mine tag relevance. However, the computation of tag similarity is strongly affected by the noisy tags in the corpus, being unable to estimate precise tag relevance. In this paper, tag refinement problem is tackled from the angle of topic modelling. Since topic model does not need explicit co-occurrence among terms (tags) in order to reach to the conclusion that they are semantically similar.

- *Knowledge/External Source Based Approaches* - Lee et al. [62, 63] derived subsumption, similarity and equivalence relations among folksonomy tags using collective intelligence of Wikipedia showing precision and recall upto 88.03 and 91.87 % respectively. Min et al. [80] identifies semantically related tags using WordNet (Disambiguation using WordNet) and Lin similarity measure. They have tested their proposed method on Flickr tags. Experiment showed that their method provides similarity improvement of 80.28 % over some other methods.

WU and ZHOU [116] viewed semantic relatedness among tags in context of the semantic relatedness among words. For this, they mentioned to use Roget's thesaurus, WordNet and Wikipedia. They also concluded that Wikipedia semantic network has a larger coverage as compared to WordNet for computing semantic relatedness.

- *Hybrid Approaches*- Uddin et al. [114] method (Mlin) for finding relationship among tags makes use of WordNet and co-occurrence metric. In addition to pair wise relationship between tag, resource and user; relationship among three is also considered. The proposed technique experimentally proved to be more effective than LCH[94], JCN[11], and LIN[20] in discovering semantic relationships among tags in Flickr and Del.icio.us dataset with F measure value of 80.28 %.
- *Other Aspects* – K.G.V.R. et al. [55] attempted to detect topics in a document by making a topic space composed of frequent document combinations that have common set of keywords, showing these keywords representing the same topic.

It is important to note that semantic similarity is different from semantic relatedness, as the later covers concepts such as antonymy and meronymy. However, it is observed that these terms are used interchangeably. In essence, semantic similarity and semantic relatedness mean, "How much does term X has to do with term Y?" [116]. There are many ways for estimating semantic similarity such as by finding distance between the words as proposed in [73, 96]. The outcome distance is more often represented as a number between 0 and 1, where 1 stands for extreme high similarity/relatedness, and 0 means little-to-none [115]. Moreover, the results of each approach are different. For example the strength or weight of similarity involving two tags based on two different measures could be dHierarchical /Tag taxonomyifferent. Similarity between two tags based on WordNet could be changed from similarity based on cosine measure [1].

3.2 Co-occurring tags

Tags co-occur in a variety of ways as identified by Halpin and shepard in [45]. In *Super-Class Relationship* tags that co-occur often represent general to specific relationship for example, '*music*' co-occurs with both '*piano*' and '*guitar*', and can be taken as super-class of both. In comparison, '*piano*' most likely does not co-occur with, more likely tags other than '*music*' and generally co-occurs with '*music*' so it is possible for it to be subclass of '*music'*. In *Facet Relationships* tags that co-occur often might have structured or facet relationship. These may be dyads or triads. For example, '*book*' and *'author* and '*Mark Twain*' is a triadic ('*triple*' in Semantic Web) relationship, and if these co-occur quite often they are most likely a facet. In fact, one would expect that most co-occurrences are dyads, like '*author*' and '*Zadie Smith*', or '*book*' and '*Mark Twain*'.

According to simpson [101] co-occurrence between tags takes place when both tags are used with the identical resource. Co-occurrence can be inner or outer. In *Inner co-occurrence,* a single user applies both tags to a resource and in *outer co-occurrence*, both tags are assigned by different users to a resource [67]. Let's have a look at the approaches.

- *Statistical / Mathematical Approaches* – Simple co-occurrence of two tags $tag1$ and $tag2$ is calculated by simply counting the number of resources (Urls, photos etc.) that are labelled with both $tag1$ and $tag2$ as illustrated in Eq 6.

$$co{-}occurrence(tag1, tag2) = |tag1 \cup tag2| \tag{6}$$

However, the drawback of this simple co-occurrence is that it gives more weightage to pairs of tags whose occurrence is very frequent. As a result, frequent tags will co-occur

more often than infrequent tags even if they are not related. This problem can be solved by Normalization. There are two types of Normalization *Symmetric* and *Asymmetric* [100].

For *Symmetric normalization* Abbasi et al. [2] compared cosine similarity, Jaccard coefficient, Dice as defined in EQ 1, 2 and 3 in section 3.1.1. They concluded that the Dice co-efficient gives higher value to co-occurring tags than the Jaccard co-efficient. Secondly, Jaccard co-efficient penalizes tags which do not co-occur very often. Zhang et al. [131] utilized Mutual information (Minfo) for symmetric co-occurrence as shown in EQ 7. The low value of Minfo indicates the two tags never co-occur, in contrast high value means high correlation.

$$Minfo\,(tag1,\ tag2) = \frac{\log(co{-}occurrence(tag1,\ tag2))}{occurrence(tag1).occurrence(tag2)} \tag{7}$$

Normalization in Asymmetric takes place by using frequency of one of the tags [65, 100] as shown in Eq 8.

$$P(tag2|tag1) = \frac{co{-}occurrence(tag1\,,tag2)}{occurrence(tag1)} \tag{8}$$

Sigurbjörnsson et al. [100] concluded that according to experiments Jaccard symmetric coefficient is good in discovering equivalent tags. In comparison, Asymmetric tag co-occurrence is able to provide a more diverse candidate tags to annotate a resource.

Wu et al. [115] studied user vocabulary of tags. They linked folksonomy tags based on collaborative tagging from users using co-occurring tags, users and resources to form a semantically connected network of folksonomy. Fujimura et al. [36] proposed dimensional placement of tags in the tag cloud according to their co-occurrence facilitating tag search in large scale tag clouds. Their approach does not overlap tags in the cloud. Through k-dense they computed centrality of tags and assigned them height accordingly. In this way, relevant resources can be found even if they don't exist in immediate neighbours of a tag. Freq / FolkRank algorithm shows bias towards high-frequency tags, i.e. to hyperonyms [20].

Tibely et al. [109] focussed on the statistical properties of tag occurrence in tagged networks with the help of 2D tag distance distribution for the relative positions in the DAG (directed acyclic graph). Fig. 2 is the diagrammatical representation of the scheme. The

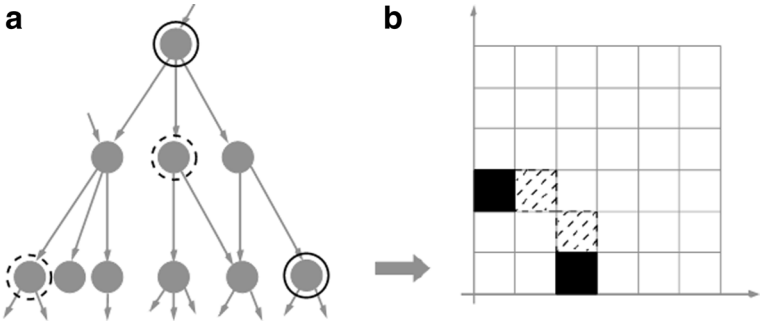


Fig. 2 **a** A small piece of a DAG with two pairs of tags are chosen, solid filled circles represents ancestor–descendant relation, whereas dashed circles represent 'uncle–nephew' pair. **b** In parallel cells of the tag-distance distribution are displayed in solid black colour and with dashed lines, respectively [109]

DAG of hierarchy between the tags is already defined. First column of cells and the bottom row contains co-occurring pairs of tags which are in direct ancestor–descendant relation, whereas the diagonal cells keep up a correspondence to pairs in which the two tags are similarly deep in diverse branches from their lowest common ancestor.

- *Knowledge/External Source Based Approaches*- Garcia et al. [37] disambiguate tags using DBpedia and TSR (TAGora sense repository[4]). They have built index and in a triple, stored the title, term frequencies, disambiguation, number of incoming links, and redirection links of wiki articles. Titles are stored in different forms like in lowercase letters, concatenated title. When TSR is queried for a tag, it returns all the DBpedia resources representing different senses of the tag along with weight given to each tag and term frequencies for each of the wiki resource. They consider co-occurring tags for a resource (as the context for any of those tags in folksonomy) and senses in vector representation. For any tag, their sense and context vectors are compared through cosine similarity. But their presented approach is non-experimented. They just tested the algorithm on tags from real data.

  In Flickr site, there are Flickr clusters for disambiguation of tag. These clusters have tags based on their co-occurrence. But the drawback of this approach is that synonyms are not clustered and if a resource is assigned a tag that does not co-occur with other related tags, that image will not appear in user's search results even if he/she searches for that tag.

  Lee et al. [66] proposed a system *tagplus* that uses homonyms and synonyms from WordNet to retrieve more relevant images from Flickr. They make use of synset *id* present in WordNet. Due to no homonym control in this approach, Flickr may return images that are not relevant to user entered keyword or sense even if he or she uses highly relevant tag. But it reduces synonymy problem by searching for synonyms of the user-entered keywords.

  Tag sense disambiguation (TSD) is experimented on vocabulary of social tags, thereby enabling users to know the sense of each tag with the help of Wikipedia. To discover the accurate mappings from Del.icio.us tags to Wikipedia articles, Local eighbor tags, the Global eighbor tags, and finally the Eighbor tags have been utilized. These useful keywords play useful role in disambiguating the sense of each tag based on the tag co-occurrences. The main objective of TSD is that the sense or meaning of a tag can be disambiguated by the help of its neighbour tags, which acts as a context. Neighbour tags can be defined as the tags that co-occur very frequently with the tag. The underlying principle behind this co-occurrence-based approach is that the frequent co-occurrences of two tags can be taken as they have high semantic relatedness among them. This approach is based on the collective intelligence hidden in folksonomies [64].

  The drawback identified by [10] regarding co-occurrence is that tag co-occurrence is not a solution of homonymy/polysemy problem when used alone.

3.3 Clustering

Folksonomies have nested groups of tags associated to common topics [101]. Clustering in folksonomy can be viewed as clusters of tags, context dependent clusters of tags, clusters of resources, clusters of users or combination of them. Let's have a look at the approaches.

[4] http://www.tagora-project.eu/

- *Statistical/ Mathematical Approaches*– Clustering techniques keeping in view only tagging information and tag co-occurrence to find out semantically related sets of tags and resources, out of folksonomy, are achieved in [12] Flickr clusters[5]. Such techniques require only statistical analysis tags and they lack semantic information. As a result, they quite frequently yield clusters of co-occurring tags, which can neither be mapped to an actual topic nor understood by a user. Moreover, most of the time these clusters are unable to solve the problem of tag synonymy, the reason is synonymous tags are usually given by users from diverse background and they rarely co-occur [40].

  Agglomerative clustering algorithm, Asymmetric hierarchical clustering, Hierarchical divisive clustering algorithms, Probabilistic Latent Semantic Indexing (PLSI) and User-Categorize tag (UCTag) have been tested and proposed by [4, 33, 46, 101] respectively in order to make tag clusters. However, some clusters produced may be too large if utilized for navigation and for that, removing unpopular tags before clustering can be useful. Hierarchical agglomerative clustering of tags also proved to be effective in personalized navigational recommendations. However, choice of cluster selection can further improve the recommendations by deleting clusters which are not directly linked to the user's query [1, 86].Clusters of tags can be successfully utilized in order to find out both the user's interest as well as topic of a resource [29].

  A co-clustering approach is proposed in [67] to yield clusters containing both resources and user annotation (tags). The technique makes use of groups of correlated tags and social data sources. It also considers the semantics in addition to the social aspect of resources accompanying tags in a reasonable way. Cluster of tag, resource and user, simultaneously using centroid based approach achieved by cosine similarity is proposed in [72].

  Among these approaches, Agglomerative Clustering algorithm has been used in most of the recent researches because it is quite flexible and can make required number of levels and cluster sizes. However, Xu et al. [120] argued that use of K-means or Hierarchical Agglomerative Clustering techniques for making tag clusters work well if tags are scattered spherically and evenly in data space. These techniques will not be effective if distribution is arbitrary, for example "S" shape. As freedom of tagging inhibits any surety of distribution of tags evenly or spherically, they proposed tag clustering based on kernel information propagation via random walk on graph to resolve this issue. They did experiments on six datasets and compared results of this clustering technique with others.
- *Knowledge/External Source Based Approaches* –Mirizzi et al. [84] states in their work that Wikipedia categories that help clustering wiki articles are reflected in DBpedia (cluster resource sets). All DBpedia categories are skos:concept. But the documents are associated with categories they specifically belong to. They are not associated with each and every category which they belong to in some manner. Haridas et al. [47] says that if clustering is done using a discrete knowledge base, clusters don't show information or semantics about the concept. In DMOZ (Directory Mozilla) resembling hierarchies, it is easy to present different semantic relations other than just subsumption.
- *Hybrid Approaches*- Lu et al. [72] clustered simultaneously the users, tags and resources as these three are interrelated in tripartite structure of folksonomy. They calculated random clusters centroid based on user, tag and resource vectors and then included these three nodes in a cluster having least distance in cosine similarity with the centroid. In this clustering approach, contents of a web resource are not considered as compared to k-

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

means clustering algorithm that uses word vectors. So this method can be implemented on different types of web resources like video, images etc. But as they compared the tripartite link structure only, false associations among tags and resource cannot be identified. They used DMOZ in order to validate resource clusters extracted from tripartite structure of folksonomy. SEMSOC [40] framework (SEMantic, SOcial, Content similarity) suggested clustering process of multimedia resources. It makes use of jointly semantic, social and content-based information, however, were based mostly on tag co-occurrences.

- *Other Aspects* – K.G.V.R. et al. [55] proposed document clustering by means of a hierarchical algorithm and using Wikipedia as an external knowledge source. They first mine frequent itemsets (sets of words that occur frequently and can be used for making clusters) for topic detection within a document and clustering of that document with other documents. First, tf-idf scores are assigned to each document in a cluster. Then Wikipedia categories and outlinks are used. Each cluster is labelled belonging to relevant Wikipedia categories (whose occurrence frequency is top k for all documents in a cluster). Their evaluation was based on five standard datasets and they claimed that their results outperformed the current state of the art methods.

3.4 Hierarchical /Tag taxonomy

Hierarchy is considered a classical semantic relationship. This section is all about the approaches that bring hierarchical structure to folksonomy.

- *Statistical/Mathematical Approaches* – Aras et al. [4] presented a tag cloud in which tags can be explored at different hierarchy levels, which gave increased semantic density and focused result. They have used cosine similarity for normalized tag co-occurrence and also considered term context. Agglomerative clustering algorithm has been utilized. Evaluation showed that the users were more satisfied with Semantic Cloud user interface than the standard user interface of folksonomy (in this case Del.icio.us).

  Search result classification based on hierarchical clusters (c-clustering) and zoom based navigation is proposed by Rástočný et al. [95] to improve web search results. Hierarchical clusters make use of semantic properties of search results to produce clusters and hence do not require to be predefined by domain specialists. It also solves the navigational pitfalls of faceted browsing.

  Eda et al. [33] used folksonomy triples to organize tags in generalized and specialized relationships. Using Probabilistic Latent Semantic Indexing (PLSI), they distinguished between subjective and objective tags and then arranged the objective ones into a hierarchy in a Directed Acyclic Graph. They measured the subjectivity of a tag by computing its entropy.

  Considering the different levels or degrees of tag generality (or tag abstractness), for highlighting hierarchical relationships that exist among concepts, [14] suggested by their results that centrality and entropy measures can distinguish well between abstract and concrete terms. Moreover, the tag co-occurrence graph is a key important input to centrality measures as against to using tag similarity graphs to compute abstractness. Tag generality vs. popularity problem is also taken into account and it is concluded that, in fact, popularity seems to be a fairly good indication of the true generality of a particular tag.

  The approach used in [102] is based on the conclusion that co-tags are appropriate for developing ontological structures based on folksonomies. Cosine similarity among tag

vectors is also an appropriate tool to identify alike tags. An unsupervised method for generating such structure taking into account combination of association rule mining and the underlying tagged material has been utilized for generating a semantic representation of each tag. The semantic depiction of the tags is an essential component of the structure generated.

Daud et al. [27] presented ontology of folksonomy taking into account users, tags and resources all at the same time. They named their proposed approach as Actor-Concept-Instance-Topic (ACIT). Their approach outperforms User-Word-Topic (UWT) and Tag-Topic (TT) approaches in accuracy by 8.4 % and 7.4 % respectively.

Tang et al. [108] formalized a novel problem of ontology learning from folksonomies. By taking into consideration, a probabilistic topic model to represent the tags and their annotated documents, they proposed four divergence measures (Tag, Hypernym, Merging, and Keep). This algorithm is utilized to construct a hierarchical structure from tags. Results of experiment conducted on two different types of real-world datasets prove to be effective in learning the ontological hierarchy from social tags.

Kawakubo et al. [54] introduced hierarchical relation by computing visual, text-based and combined concept vectors. First, they calculated entropy and JS divergence for these three vectors. Degree of relatedness among the concept vectors has been analysed and hierarchical relations among tags have been extracted. They constructed three different ontologies, each of them based on one of the concept vectors, among which they found that the one based on combined features is better than the other two. The noise removal accuracy rate on the average for selected images was 92 % and for randomly selected images was 70 %.

- *Knowledge/External Source Based Approaches*- YAGO project [121] worked on structured information extracted from Wikipedia. It makes use of Wikipedia category system and redirects and considers fourteen types of relations. But it does not completely make use of the hierarchy provided by Wikipedia category system. It just maps end points of categories to WordNet hierarchy. FreeBase project5 also attempts to make an online accessible data base that can be edited as a wiki.

  Kobilarov et al. [58] mentioned in his work, that DBpedia entities have been arranged in four different hierarchies: SKOS representation of Wikipedia categories, DBpedia hierarchy YAGO ontology, UMBEL ontology and DBpedia hierarchy (developed manually).

  Tomuro et al. [110] built ontology from folksonomic tags. Using Domain Similarity Clustering by Committee (DSCBC) algorithm, they made clusters of related tags using Wikipedia knowledge source. In these committees, ambiguous tags are included in each related cluster based on relevance to show their different senses and then ontology from these disambiguated tags using agglomerative clustering algorithm is generated.
- *Folk based approaches*- Structured folksonomies with predefined structure (e.g. hierarchical) have some pitfalls (1) restriction on tagging because of limited pre- defined vocabulary and (2) Selection of tags, which is time consuming manual effort. Yoo et al. [126] proposed a technique based on the idea that when a user enters the tag, he/she must also define its category. This tag is called categorized tag (CT). CTs are added to collaborative structured folksonomy(CSF) showing tag category relation supported by most of the users. A CT based organizational layer is built on top of CSF for organizational knowledge classification and enables users to find appropriate knowledge. Authors compared their technique with flat folksonomy and claimed to be effective in retrieval.

  Yoo and Suh. [127] proposed a prototype User-Categorized Tag (UCTag) in the form of a document management system. In this system users can assign tags and specify their category as well. Thus, a structured folksonomy based on user's consensus emerges in

which tags are included in different categories. But the relationships in this hierarchy correspond only to '*has-a-* relationship' type.

Ding et al. [30] proposed upper tag ontology based on tagging behaviour. Mika et al. [78] added to the folksonomy the user's aspect by introducing Actor-Concept-Instance model.

- *Heuristics Based Approaches* -In [113], an approach based on heuristic regulations and deep syntactic analysis for taxonomy construction has been utilized. In the first step, tags are obtained from the tag clouds of domain folksonomy websites. The folksonomy tags play role of target domain taxonomy. The taxonomy is constructed without human intervention based on heuristic principles and deep syntactic analysis. Heuristic rules approach traditionally has the trait of relatively low recall but high precision rate. In comparison, deep syntactic analysis has a higher recall however lower precision rate. Two algorithms have been combined applying heuristic rules analysis first and then a concept–relationship acquirement algorithm to steer clear of the low recall. But the challenge is heuristic patterns are uncommon to be discovered in tags.
- *Other Aspects* - Pirrone et al. [91] took text of wiki articles into analysis to derive relationships and concepts. They extracted relationships from both Wikipedia link structure and text. After information extraction, contents are structured in the form of ontology. They have used table of contents in the wiki pages for extracting semantic relations. In their proposed methodology, semantic sense extraction is done using table of content tree and text of the section. Sense of a section is extracted by comparing domain ontology with the table of contents. The link analysis is a good source of relating terms to each other.

  In recent times, numbers of researches are being done on integration of folksonomic and ontological approach. Hierarchical ontology development based on existing hierarchies like DMOZ gives better results instead of making hierarchies from the scratch. However, by using knowledge sources like WordNet, AWS, IMDB etc., resulting hierarchy is a binary tree and the clusters do not show information or semantic about the concept of the child nodes [47].Chen et al. [23] says that as it is a cumbersome job for domain experts to make an ontology from scratch, so folksonomy is a very good knowledge source to build ontology that will also reflect collective intelligence.

3.5 Tag-pairs subsumption

Subsumption relation between any tag tag_x and tag_y can be defined as in [99], tag tag_x subsumes tag tag_y, means that everywhere when the tag tag_y is used, tag_x can also be used without ambiguity. The subsumption relation between tag tag_x and tag_y is represented as given in Eq 9.

$$tag_x \rightarrow tag_y \tag{9}$$

Subsumption relation is directional, that is, $tag_x \rightarrow tag_y$ does not mean $tag_y \rightarrow tag_x$. But the subsumption has transitivity property, that is $tag_x \rightarrow tag_y$ and $tag_y \rightarrow tag_z$ means $tag_x \rightarrow tag_z$. Subsumption relation is stricter than similarity metric. Now let's have a look at the approaches.

- *Statistical/Mathematical Approaches* – Han et al. [46] makes use of Asymmetric hierarchical clustering algorithm to find tag subsumptions. They have used tag co-occurrence to measure similarity among cluster tags and dissimilarity among different clusters. Resulting hierarchy reflects knowledge of the users. Mo et al. [85] utilized entropy to measure tag-pairs subsumption relationships in diigo and Del.icio.us.

Si et al. [99] in his work proposed TAG-TAG, TAG-WORD and TAG-REASON. The last two give weightage to the content of document to help estimation. The results showed that the proposed methods performed better than the similarity-based hierarchical clustering in order to dig out subsumption relations.

- *Hybrid Approaches*- Lee et al. [63] FolksoViz, a statistical representation for digging out subsumption relationships keeping in view the number of occurrence of each tag in the Wikipedia texts, along with using the TSD (Tag Sense Disambiguation) technique for mapping each tag to an equivalent Wikipedia text. The derived subsumption pairs are shown successfully on the display screen. The experiment shows that the FolksoViz manages to dig out the right subsumption pairs precisely.

3.6 Some other semantic relationships

- *Non Taxonomic relation discovery* - Non-taxonomic refers to absence of hierarchy among the classes. Taxonomic relations such as *subclass*, *superclass*, *is-a* or *has-a* are lacking in non-taxonomic relations. For example '*Polio affects children*'. Classes will be '*Polio*' and '*children*' and the relation between them is '*affects*'. In general, two tasks have to be performed for non-taxonomic relationships. First is to find out which concepts are correlated. Secondly, it is required to dig out how these concepts are linked, so that the name can be given to the relationship [111]. Trabelsi et al. [111] worked on discovery of non-taxonomic relation in folksonomy. In their work triadic concepts have been used in order to find out and select related tags. External sources (Wikipedia, WordNet and Google) are utilized for tags filtering and non-taxonomic relationships discovery.
- *Bursty Tags*- Yao et al. [123] identified bursty tags and events from the folksonomy tags. They make use of temporal information for burst detection. They extracted temporal tag graphs from the tag space by dividing tag space into time intervals based on tags time stamps. These temporal tag graphs are much smaller in size than the whole tag space and maintain only those tags and their correlations that have some bursty information. From these local tag graphs, they identified bursty tags and edges using a generative Gaussian distribution and Probabilistic model.
- *Time and Location Tags*- Baba et al. [9] not only worked on finding the time and/or location related tags on flicker, but also extracted the concepts related to a tag in a machine-understandable way. Another work in this direction by Zhang et al. [132], computes connection or relationship among tags by analysing their distributions over time and space. In other words, their work is based on digging out tags with similar geographic and temporal patterns of use. Using a dataset obtained from Flickr, Flickr photo tags are clustered based on their geographic and temporal patterns.
- *Tagging Motivation/Self intention based*- Strohmaier et al. [104] highlighted different tagging motivations and concluded that motivation behind tagging effects tagging behaviour of the users (selection of tags) in folksonomy. Making these motivations as basis, users and tags can be categorized. Cantador et al. [19] proposed classification of tags into purpose-oriented categories namely context or content based tags. By purpose-oriented they mean to categorize according to their intentions. Semantics of these categories of tags have been retrieved from Wikipedia and WordNet. The results have significant accuracy. Körner et al. [59] identified various quantitative measures (Tag/Resource Ratio, Orphaned Tag Ratio, Conditional Tag Entropy, Overlap Factor

and Tag/Title Intersection Ratio) to identify Categorizer and Describer users based on their tagging behaviour. Categorizers use tags for categorization of resources while Describers use tags for description of resources. All measures they identified work well but are not equally useful. Among all these measures tag/Resource ratio prove to be the best.

4 Protection of folksonomy structure

Instead of focusing just on the tag, resource and user association discovery, we also need to consider protection of valid relationships. By this consideration we mean to handle issue of spam tags and spam users. In this section, we are focussing on techniques covering this aspect, so that folksonomy maintains its correct structure with time.

Tagging systems are quite easy and cheap target for spammers as compared to spamming through online advertising, email systems and search engines. User can add any content, generate spam annotations anonymously without any cost. Tag collision [70], where people either purposely or unintentionally use the same tags, for equally valid yet not related contents. The intention for making false associations among tags and resources can be, for example, by assigning tags that are popular bring their resources higher in search result ranking. Apparently, no one is harmed by spam tags on web but good web information resources become difficult to be found among all the content.

Spam can be introduced at resource level, in the form of spam posts (incorrect Tag-To-Resource association) or through spam user accounts. Hayati et al. [49] presented a survey and evaluation of anti-spam methods in Web 2.0. They evaluated the methods based on whether they used a *preventive strategy* or a *detective one*. Authors of [31, 92] classified anti-spam techniques as *Prevention*, *Detection* and *Demotion* based. Spam detection/prevention approaches can also be classified on level basis that is user level or post level. Post level means that individual posts are marked as spam or otherwise, whereas user level means all or none of the posts of a user is marked as spam.

4.1 Spam posts/Tag spam

Spam post means incorrect tag to resource association. Misleading tags that are generated in order to boost the visibility of some resources or minimally to confuse and mislead the users. Let's have a look at the approaches.

- *Statistical / Mathematical Approaches*- Combining KNN algorithm with tag clustering to filter noisy tags is proposed by Pan et al.[89]. By doing so they improved the accuracy of recommendations. The precision results of this technique for the M-Eco and Moivelens dataset are 73.9 % and 87.1 % respectively in comparison with *TagNeighbor with Clustering*, *TagNeighbor*, *Collaborative Filtering* and the *Pure Tag* techniques.
- *Folk Based Approaches*- The performance of the algorithms based on static user data analysis has been presented in many studies in order to combat with tag spam, but either they do not give precise evaluation or the algorithms' performances are not appreciably good. Liu et al. [69] makes use of dynamic user behaviour data for the notion that users' behaviours in social tagging system can mirror the quality of tags more precisely. By making different categories of participants' behaviours, tag-associated actions are extracted to estimate whether tag is spam or not, and then proposed algorithm filters the tag spam as an outcome of social search. The observed results demonstrate that method indeed

outperforms the already present methods based on static data and successfully defends against the tag spam in a variety of spam attacks.

Zhai et al. [130] proposed a technique in which personalized experience is assigned by a user to other annotators using correlation. This results in a ranked list, according to his personalized experience with other annotators. For those annotators who don't have common tags with other users, socially enhanced mechanism is used to link users by some references. For evaluation they compared efficiency of SpamClean model to the occurrence, coincidence and boolean model, on different threats like collusive, normal and tricky attacks. SpamClean effectively defends against spam tag.

Koutrika et al. [60] assigned relevance numbers to web resources based on the number of common tags they share. This is a language-independent method. Krause et al. [61] identified spam in their work on post level so that only malicious posts are blocked and not the rest by any user. They outlined four feature set categories to tackle spam and evaluated them against machine- learning techniques.

- *Other Aspects*- Yhang et al. [122] proposed method is based on text mining approach which could find out the relationships between web pages and also among tags. In the first step, Web pages and their tags are clustered using self-organizing map algorithm. A labelling process is applied on the trained map to find out the relationships between web pages and among tags. The detection of spam tag could then be achieved by looking at the semantic relatedness between a tag and its tagged web page.

  Zhai [129] proposed spam-proof tagging system leads to a good quality tag search. The proposed technique is based on four key factors including demotion-based strategy, reputation, altruistic users and social networking. The proposed technique, upgrades/ degrades the ranks of correct/incorrect content items in the search results by taking into account personalized users' reliability degrees and responsible users. Thus preventing clients from picking unwanted contents.

4.2 Spam users /Social spam

In [57], authors identified users that created semantic noise in the folksonomy. They showed it experimentally that hyperactive taggers perform more tagging actions comprising 40 % of all. Hence, removing these users reduces semantic noise from folksonomy. The techniques adopted for spam users are mostly tested at both user and post level. Let's have a look at the approaches.

- *Hybrid Approaches*- Markines et al. [75] addressed different properties of spam in social tagging systems to differentiate spam users from legitimate ones. According to the author removing spam at post level is most appropriate. Among the six features they have used to identify spam, three are at resource level, two at post level and one for identifying spam users.

  Based on work of [106], that is scoring and semantic analysis of tags using tag score shows 95.0 % performance. Performance is further improved to 96.8 % when selective evaluation using the white tag and black tag concepts has been used. Tag scoring seems to be powerful method for discriminating spammers, but when a spammer uses popular tags to cover-up as a legitimate user, detection becomes difficult. To deal with these drawbacks of tag score, features using semantic similarity are implemented. When semantic attributes are united with the tag features the precision increases from 96.8 to 98.0 %. In experiments comparing the feature performance at the post level and the user level, the performance of the user level was slightly better.

Poorgholami et al. [92] considered tags, resources, users and relations among them and highlighted set of features (Tag spamicity, Legal and illegal domain, coincidence and Network features). In their work they claimed that above mentioned features are effective in detection of spammers. The reliability of presented features is over 95 %, and combination of them is 99 %.These features are used for various machine learning algorithms to sort out spammers and achieve 99 % accuracy.

5 Enrich query and search results

Structured folksonomy enables elicitation of precise search results. In addition, *mapping query keywords for disambiguation and semantic clarity*, *ranking search results*, *secondary tags* and *multilingualism* also significantly improves precision in search results. This section is planned to focus on these aspects.

5.1 Mapping and ranking

Search engine results use only lexical information and web page importance on web to rank results. Folksonomies are difficult to navigate if tags are presented as long lists [101]. Now let's have a look at the approaches.

- *Statistical/Mathematical Approaches*- Chen et al. [24] argued that WordNet is too fine (many tags in folksonomy match to one sense not to all the senses available in the WordNet of a particular word) as well as too coarse (does not cover senses of a word in all domains) in defining granularity of word senses. Therefore, it is not fit for social tagging system. A technique based on non-negative matrix factorization (NMF) is proposed for automatic discovery of topic sense from tags and then used for tag disambiguation. The aim of the technique is to achieve precision in searching of resources.

  A technique for providing users with more specific keywords to replace or enhance the meaning of abstract tags when giving query and to precise the search is proposed by Xia et al. [117]. In the first step, ontology in which concepts are categorized in three semantic levels (General, basic and specific) to detect abstract tags is built. To confirm wether the selected tags in the first step are abstract or not and also to identify specific tags they utilize co-occurrence for tag context and K-NN with Gaussian weight for image context of a tag in the second step. For image context, similarity of both visual and textual features are combined because author mentioned that it gives 8 % more improvement in detecting abstract tags as compared to using visual and textual similarity individually. In addition to identify specific tags in the second step, all the in-between nodes between abstract tags and specific tags from the ontology developed in the first step are added to provide set of concrete tags for abstract tags.
- *Knowledge/External Source Based Approaches*- Mirizzi et al. [81, 84] presented a tool *Not Only Tag*[6] by mapping keywords to DBpedia resources and by using DBpedias's ontological structure to enrich its meaning showing results in the form of a tag cloud. It ranks resources using a hybrid ranking algorithm. Resources are ranked based on their relevance with the query and other related connected nodes in DBpedia graph, rather than calculating

[6] http://sisinflab.poliba.it/not-only-tag/

individual resource importance separately as done in PageRank algorithm.

The DBpediaRanker algorithm computes relevance among DBpedia nodes by exploiting link structure, title and abstract comparisons, by querying social bookmarking systems as well as by considering web search engine results as shown in Fig. 3 [82].This algorithm has statistically significant results over the other algorithm with which it was compared. The same authors presented LED [82] (Lookup Discover Explore) to provide exploratory search using their RDF ranker in DBpedia. They say if users are helped by semantic tags, they can save monthly 10 min of each user and thus, in aggregate will save 4.1 million working hours yearly. They find relevant resources by discovering them in the neighbourhood of a resource node.

Lin et al. [68] attempted to combine ontologies and folksonomies to improve search and navigation. Bindelli et al. [15] presented TagOnto system that performs mapping of folksonomy to ontology providing users access to folksonomy system with search and navigation features that are peculiar to ontologies. Passant [90] attempted to combine weblogs and ontology for better information retrieval by mapping folksonomy tags to domain ontology. He used SIOC ontology.

Ronzano et al. [97] said in their paper that if web resources are characterized according to the concept they represent instead of keywords, it may increase precision. They proposed *Tagpedia*, a general-domain encyclopedia of tags to provide web content descriptions through Wikipedia. It covers 84 % of the considered tags. They integrated this semantic resource into SemKey [74]. When user selects a tag, he or she can further select the sense of the tag. A weak point of this approach is that Tagpedia does not provide coverage on non conventional tags and there are no semantic relations defined among the Syntag sets. Furthermore, the same Syntag sets are not available in multiple languages,

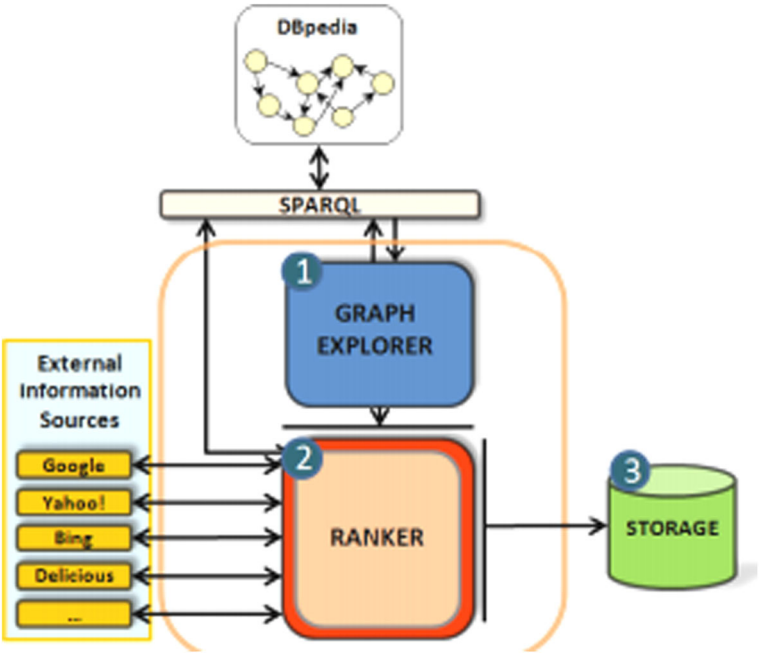


Fig. 3 DBpedia Ranker [82]

thus there is no support for multilingualism. Hence, search in Tagpedia lacks multilingualism and support for non conventional tags.

Iijima et al. [50] proposed linked Flickr search, by integrating DBpedia, user preference data and folksonomy tags. When the user enters a query, the tag is searched on DBpedia and all classes the tag belongs to are returned. These classes or class instances are then ranked according to their weights from the user's search logs. Flickr images are searched by giving initial tag entered by the user and the DBpedia instance the user selects. Results are evaluated by comparing it with Flickr Wrapper [35]. But evaluation results show that precision and recall values are lower than Flickr Wrapper with increased unexpectedness.

Dellschaft et al. [28] presented sensible search by querying TAGora Sense Repository to give senses list for a tag after normalizing it and assign weight to them according to their importance. It retrieves different senses using DBpedia:hasDBpediaSenseInfo property. Mirizzi et al. [83] further presented Semantic Wonder Cloud that supports exploratory search using the same hybrid approach as described above and gave statistically significant results. They provided exploratory search as O'Brien wrote in an article [88] that: "*The Web, they say, is leaving the era of search and entering one of discovery. What's the difference? Search is what you do when you're looking for something. Discovery is when something wonderful that you didn't know existed, or didn't know how to ask for, finds you*."

Choudhury et al. [25] in their work, semantically enriched tag cloud of YouTube by linking it with the Linked Data Cloud and expanding and ranking the tag space. For semantic enrichment of tag space, they used their own dataset to generate related videos based on temporal, textual, geospatial and social context. Then, they further expanded it by tag co-occurrence analysis. However, they have not fully implemented Tag-To-Concept mapping module. Quality of their tag enhancement and quality of ranking was upto 80 % accurate. Similarly, the tag enrichment process when evaluated showed that content understanding is improved.

Stampouli et al. [103] dealt with tag disambiguation and improved content retrieval quality in Flickr using mashup. They showed through a case study that this system provides high retrieval quality. Figure 4 shows a graphical representation of the system.

Chandramouli et al. [22] presented Semantic Concept mapping that leads to Hypernym Discovery (SCMTHD) algorithms resulting in accuracy improvement from 49 % (single-user environment) to 58 % (collaborative environment).They mapped tags to synsets of WordNet to get semantic concepts. But for semantic concept mapping they do not consider the problem of ambiguity. THD uses online resource for hypernym discovery. They used Wikipedia to increase entity coverage.

- *Other Aspects*- Cucerzan et al. [26] identified named entities and disambiguated those using Wikipedia data. The accuracy of identifying named entities from within the text was 91.4 %. Technique for finding temporal semantic context of a concept (associated words, context graph, associated concepts, context communities and example sentences) that can be effectively used for query suggestions, faceted searching and trend analysis is proposed by Xu et al. [118]. They claimed that proposed technique helps in discovering semantic context automatically as compared to manualy generated context repository. The technique is tested for the effectiveness and accuracy.

5.2 Secondary tags

There is an information overload on the web. However, meaningful metadata can increase precision substantially. Searching solely based on user's generated tags is not efficient due to

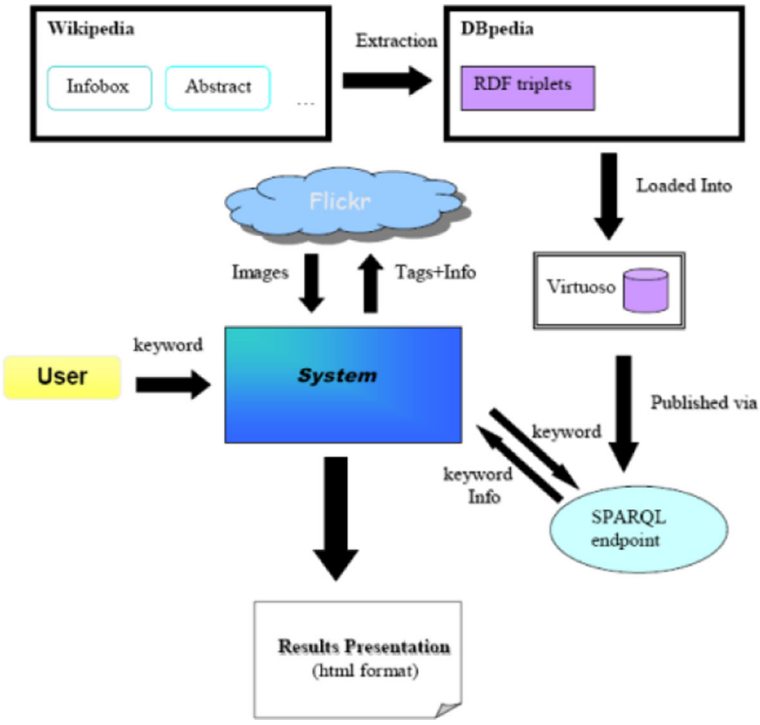


Fig. 4 Proposed frameworks for Tag disambiguation by Stampouli et al. [103]

variety of reasons. The three main reasons are: firstly, the usual number of tags assigned to a document is from 0 to 19 but among them mostly just 2 is the modal number. Secondly, due to presence of like polysemy, synonymy etc. Third, in some cases tags may not represent true metadata. e.g., if a user tags a resource with his/her opinion about the resource like'interesting', then this tag may not be of use for other users in searching that content/resource.

All this signifies the need for presence of some content related metadata to be added for improved retrieval of resources. If the metadata can be generated in the form of keywords that are extracted automatically and these keywords are used along with the user's created tags, result in improved search quality and precision. Let's have a look at the approaches.

- *Knowledge/External Source Based Approaches*- Awawdeh et al. [7] added metadata to user's generated tags in folksonomy using Yahoo Term Extraction API. They generated keywords from text of original document. In their previous work [75], they compared different techniques to extract terms from web documents. These techniques comprised of extracting meta tags for document description, using yahoo term extraction service and terms selection having highest term frequency. They showed through experimental results that yahoo terms added the most to the searching process. They presented Enhanced Tag Set Engine that combines yahoo terms from the document with the user's tag set.

  Faviki [79] combines tagging and Wikipedia by suggesting tags from Wikipedia concepts. But the suggested tags must be name of some Wikipedia article. The semantic

tags it provides are machine-interpretable. It makes use of Zemanta [128] API for semantic tag suggestion. Zemanta is basically a blogging plug-in for firefox and can suggest tags from Wikipedia and user content.

Table 1 Feature Set Summary

Aspects	Approaches	Feature Set
Semantics	Wikipedia	Wiki Text (label, abstract, Wiki page sections) Link Structure Named Entity Recognition Mapping tags to Wiki Articles(subsumption,Simmilarity) Term Disambiguation Page in Wikipedia Semantic relations among tags Semantic Concept Mapping Use in ontology construction
	DBpedia	Mashups integrating DBpedia and folksonomy data YAGO TAGora Sense Repository Hybird Approaches AGora sense repository (DBpedia:hasDBpediaSenseInfo) dbpprop:disambiguates Tag mapping to Linked Data Cloud Faviki
	WordNet	Flickr Clusters Integration of WordNet and Folksonomy Grounding tags to WordNet synsets Lin Similarity Measure Hypernym Discovery and Synonymy problem
	Ontology	Folksonomy ontologies Ontology construction from folksonomy Ontological enrichment of tags meaning
	Statistical	Co-Occurrence Cosine Similarity Lin,Mlin Similarity Jaccard,dice,Mutaul information Google Normalised distance AJSD Semantic relatedness measures Mutual contextualization of tags, users and resources Emerging semantic from folks pragmatic behaviour
	Folk	Folk based categorization (CT, UcTag)
Tag-to-Resource Association	Spam Post Spam Users	Resource and concept matching (tag recommendation from wikipedia) Post level Spam Detection
Multilingualism	Wikipedia	Multilingual Wikipedia
	DBpedia	Titles and abstracts and infoboxes of Wiki articles available in multiple languages
	WordNet	Multilingual WordNet
	DMOZ	Hierarchical Ontology Interest Hierarchy Construction Resource Mapping
Categorization/ Classification	Wikipedia	Resource Classification Wikipedia categories Faviki (makes use of Wikipedia categories)
	Statistical Approach	Purpose oriented Tag classification Entropy Co-Occurrence Agglomerative Clustering Algorithm Hierarchical clustering of search Results
	Event Based	Time and Location Tags Self Intention Based
Search enrichment	Secondary tags	Yahoo Term Extraction API Meta tags for Resource Description Term Frequency
	Bursty tags	Bursty Tags and Bursty Events

Table 2 Feature Support Comparison

Approaches	Feature Support					
	Folksonomic/Non- Folksonomic	Non-Conventional tag coverage	Multilingualism	Disambiguation	Hierarchical clustering	Temporal
Statistical (co-occurrence and cosine)	Folksonomic	Yes	No	Yes	No	No
Folks tagging pragmatic	Folksonomic	Yes	No	Yes	No	No
Semantic Relations Through wikipedia	Folksonomic+Non- Folksonomic	Yes	Yes	Yes	Subsumption Relation	No
Tagpedia	Non- Folksonomic	No	No	Yes	No	No
Named Entities	Non- Folksonomic	Yes	Yes	Yes	No	No
Flickr Wrapper	Folksonomic	Yes	Yes	No	No	No
Tag Cloud generation Via DBpedia	Non- Folksonomic	No	Yes	Yes	Yes	No
Tagora Sense Repository	Folksonomic+Non- Folksonomic	Yes	Yes	Yes	No	No
Tag mapping to Linked Data Cloud	Folksonomic	Yes	No	Yes	No	No
Wikipedia Link Structure	Non- Folksonomic	No	Yes	Yes	Yes	No
WordNet and Flickr (lin Similarity)	Folksonomic	Yes	No	Yes	No	No
Topic Detection	Non- Folksonomic	No	Yes	Yes	Yes	No
Flickr Clusters	Folksonomic	N/A	Yes	Yes	Yes	No
Semantic Concept mapping and targeted Hypernym Discovery using WordNet	Folksonomic	No	No	Yes	No	No
SemKey	Non- Folksonomic	No	N/A	Yes	No	No
YAGO	Non- Folksonomic	No	No	Yes	Yes	No
Ontology of Folksonomy	Folksonomic	Yes	No	Yes	Yes	No
Wikipedia Categories	Non- Folksonomic	Yes	Yes	Yes	Yes	Yes
DMOZ	Non- Folksonomic	Yes	No	Yes	Yes	No
Agglomerative Clustering Algorithm	Folksonomic+Non- Folksonomic	No	No	No	Yes	No
Clustering of search results (c clustering)	Folksonomic+Non- Folksonomic	No	No	No	Yes	No
User-Categorized Tag (UCTag)	Folksonomic	Yes	Yes	Yes	Yes	No
Tri-Partite Structure	Folksonomic	Yes	No	Yes	Yes	No
Directed Acyclic Graph	Folksonomic	Yes	No	Yes (Subjective and Objective Tags)	Yes	No

- *Other Aspects*- Tan et al. [107] have referenced to papers that show that precision and recall improves by adding semantic data to XML documents. They marked up Wikipedia articles in XML form. Their approach uses semantically tagged documents to detect concepts from wiki articles using Wikipedia categories, info boxes and link structure. Precision and recall measures for the three sources show that infobox parameter name is a good source for describing the information in both; precision and recall. But a negative point is that the tag names are not implicit. In 18 different types of relations existing in WordNet, Hypernym/hyponym relation in WordNet can be used to explore words that are more specific or more general for a specific word to explore secondary tags that will increase precision.

5.3 Multilingualism support

Translating tags into different languages and utilizing them for searching makes it possible to get unexpected information in search that cannot be achieved by using only one language. Let's have look at the approaches.

- *Knowledge/External Source Based Approaches*- Wikipedia gives extensive linguistic coverage [98]. Based on this fact Gobbo [42] presented Flickrpedia, by using Wikipedia support for multilingualism. They emphasize to improve serendipity regardless of the natural language. As a result, highly unexpected and relevant photos were retrieved. However, there was no support for sense disambiguation. Among all of the applications reviewed in [18], Faviki [79] supports multilingualism by translating tags in different languages.
- *Folk based Approaches*- Jung et al. [53] support information retrieval based on multilingual tags coming from users by relating lingual practices of different folks. They translate tags into other languages to support search for multilingual resources using Google AJAX Language API.

6 Summary

The research efforts presented in this paper are summarized in Table 1 and Table 2. Table 1 is about feature set summary. In Table 2 techniques discussed in the paper have been viewed from perspective of features they support. These features include Folksonomic/Non-folksonomic, Non conventional tag coverage, Multilinguism, Disambiguation, Temporal, and Hierarchical clustering. If we look at Tagpedia, it does not provide coverage on non conventional tags. There are no relationships defined among syntag sets. The same syntag sets are not available in multiple languages. DMOZ supports Hierarchical clustering. Some of the discussed methods in the paper are not folksonomy based (Non-folksonomic) but in our opinion they can be used in social tagging model efficiently.tgroup1

7 Conclusion

Folksonomy provides a low cognitive cost system to support classification but due to its flat structure it suffers from low search precision. This paper attempts to review the different approaches for semantic incorporation in folksonomies to achieve objective of improving

precision in search and navigation. We have categorized these approaches and summarized the feature set support. Following are the concluding remarks.

Statistical approaches help to cover the vocabulary which is not present in lexical resources. However, if we compare the precision ratios of knowledge source based and statistical approaches, knowledge source based approaches perform better in disambiguation. Also, hybrid approaches that utilize features from both methods have relatively high precision than pure statistical approach.

Formal classification systems like ontologies are very good in precision but can be built for limited domains and by limited number of people-experts. Moreover, the objects to be classified in these domains are limited in number as well. To build one huge ontology from scratch that covers all domains of web resources and to update it regularly is a challenge. As far as domain ontologies are concerned, it is difficult to get consensus on domain ontologies as they are made by knowledge experts and do not have common user's consensus.

On the current web, with continuous exponential increase in the amount of content, such classification system will not be a viable solution. It needs to keep evolving to cover the emerging trends and vocabularies. Folksonomies are users driven and a non-formal way to categorize data and generate metadata while ontologies are the formal way to provide metadata for annotations. Their integration can give a very high precision. Hence, a fresh investigation in direction to integrate the folksonomic and ontological approaches can give better precision but may suffer the problem of complexity. Typical rigid taxonomies cannot tackle the challenge posed by fast evolving information space with continuous emergence of new vocabularies and trends. There may be many such terms that don't necessarily fit into some fixed set of categories. For hierarchical arrangement of tags, again external knowledge source based approaches are better with respect to precision as well as vocabulary coverage.

Lastly, bringing semantically enriched structure in folksonomy, utilizing semantics for folksonomy cleaning by removing spam posts/spam users and other aspects like multilingual, secondary tags, search query enhancement further improve precision of search results.

References

1. Abbasi RA (2010) Discovering and Exploiting Semantics in Folksonomies, Dissertation
2. Abbasi R (2011) Query expansion in folksonomies. Proceedings of the 5th international conference on Semantic and digital media technologies, pp.1–16
3. Acm yahoo grand challenge 2009 demo. [online]. available: http://www.youtube.com/watch?v=kwZsCB1tUpA.(2009)
4. Aras H, Siegel S, Malaka R (2010) Semantic cloud: An enhanced browsing interface for exploring resources in folksonomy systems, in Workshop on Visual Interfaces to the Social and Semantic Web (VISSW2010), IUI2010
5. Aschke R, Hotho A, Schmitz C, Stumme G, Ganter B (2008) Discovering Shared Conceptualizations in folksonomie. Web Semant Sci Serv Agents World Wide Web 6(1):38–53
6. Auer, Bizer C, Kobilarov G., Lehmann J, Cyganiak R, Ives Z: Dbpedia: A nucleus for a web of open data, in The Semantic Web Aberer,K., Choi,K,-S., C Noy,N., Allemang,D., Lee, K.-I.,Nixon,L., Golbeck,J., Mika,P., Maynard,D., Mizoguchi,R., Schreiber,G., Cudr-Mauroux,P (eds.),vol. 4825 of Lecture Notes in Computer Science, pp. 722–735, Springer Berlin / Heidelberg.(2007)
7. Awawdeh R, Anderson T: Improving search in tag-based systems with automatically extracted keywords, in Knowledge Science, Engineering and Management Bi,Y .,Williams, M.-A. (eds), vol. 6291 of Lecture Notes in Computer Science, pp. 378–387, Springer Berlin, Heidelberg.(2010)
8. Awawdeh R, Anderson T (2009) Improved search in tag-based systems. In Intell Syst Des Appl, ISDA'09 ninth International Conference on, pp. 288–293, IEEE
9. Baba Y, Ishikawa F, Honiden S Extracting time and location concepts related to tags. Available: http://www.km.aifb.uni-karlsruhe.de

10. Bartolini I, Patella M, Romani C (2013) SHIATSU: tagging and retrieving videos without worries. Multimed Tools Appl 63:357–385
11. Begelman G, Keller P, Smadja F (2006) Automated tag clustering: Improving search and exploration in the tag space, in Proceedings of the collaborative web tagging workshop at 15th WWW conference, pp. 15–33
12. Begelman G, Keller, P Smadja F (2006) Automated tag clustering: Improving search and exploration in the tag space, in Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, pp. 15–33
13. Benz D, Grobelnik M, Hotho A, Jäschke R, Mladenic D, Servedio VDP, Sizov S, Szomszor M (2008) Analyzing tag semantics across collaborative tagging systems, in Dagstuhl Seminar 08391–Working Group Summary
14. Benz D, Korner C, Hotho A, Stumme G, Strohmaier M (2011) One tag to bind them all: measuring term abstractness in social metadata. Springer-Verlag Berlin Heidelberg .ESWC 2011, Part II, LNCS 6644, pp. 360–374.(2011)
15. Bindelli S, Criscione C, Curino C, Drago M, Eynard D, Orsi G: Improving search and navigation by combining ontologies and social tags, in On the Move to Meaningful Internet Systems: OTM 2008 Workshops Meersman,R., Tari,Z., Herrero,P (eds), vol. 5333 of Lecture Notes in Computer Science, pp. 76–85, Springer Berlin, Heidelberg.(2008)
16. Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009) Dbpedia - a crystallization point for the web of data. J Web Sem 7(3):154–165
17. Braun S, Schmidt A, Walter A, Zacharias V (2007) The ontology maturing approach to collaborative and work-integrated ontology development: Evaluation results and future directions, in Emergent Semantics and Ontology Evolution 2007, Proceedings of the First International Workshop on Emergent Semantics and Ontology Evolution (ESOE-2007), ISWC 2007, Busan, Korea, November 12, 2007. Chen, LL, Cudr-Mauroux, P, Haase, P, Hotho, A, Ong, E (eds), vol. 292 of CEUR Workshop Proceedings, pp. 5–18
18. Braun S, Schora C, Zacharias V (2009) Semantics to the bookmarks: A review of social semantic bookmarking systems, in Semantic Systems (I-SEMANTICS 2009),5th International Conference on, Proceedings of I-KNOW 09 and I-SEMANTICS 09 A. Paschke, H. Weigand, W. Behrendt, K. Tochtermann, and T. Pellegrini, (eds), (Graz, Austria),pp. 445–454, Verlag der Technischen University Graz
19. Cantador I, Konstas I, Jose J (2011) M: Categorising social tags to improve folksonomy based recommendations. Web Semant Sci Serv Agents World Wide Web 9:1–15
20. Cattuto C, Benz, D Hotho, A Stumme G (2008) Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems, in Proceedings of the 3rd workshop on ontology learning and population (OLP3), pp. 39–43
21. Cattuto C, Benz D, Hotho A, Stumme G (2008) Semantic grounding of tag relatedness in social bookmarking systems, in proceedings of 7th International Semantic Web Conference, pp. 615–631, Springer-Verlag Berlin, Heidelberg
22. Chandramouli K, Kliegr T, Svatek V, Izquierdo E: Towards semantic tagging in collaborative environments, in Digital Signal Processing, 2009 16th International Conference on, pp. 1–6.(2009)
23. Chen WH, Cai Y, Leung FH (2010) An unsupervised method of exploring ontologies from folksonomies, in Computational Science and Its Applications (ICCSA), 2010 International Conference on, vol. 0, pp. 331–334, IEEE
24. Chen J, Feng S, Liu J (2014) Topic sense induction from social tags based on non-negative matrix factorization. Inf Sci 280:16–25
25. Choudhury S, Breslin J, Passant A (2009) Enrichment and ranking of the youtube tag space and integration with the linked data cloud, in The Semantic Web, ISWC 2009 A. Bernstein, D. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, and K. Thirunarayan, (eds), vol. 5823 of Lecture Notes in Computer Science, pp. 747–762, Springer Berlin Heidelberg
26. Cucerzan S: Large-scale named entity disambiguation based on wikipedia data, in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 708–716.(2007)
27. Daud A, Li J, Zhou L, Zhang L, Ding Y, Muhammad F (2010) Modeling ontology of folksonomy with latent semantics of tags", Web Intelligence and Intelligent Agent Technology. IEEE/WIC/ACM Int Conf on 1:516–523
28. Dellschaft K, Go¨ erlitz O, Szomszor M: Sense aware searching and exploration with my tag, in The 8th International Semantic Web Conference (ISWC 2009).(2009)
29. Di Matteo NR, Peroni S, Tamburini F, Vitali F (2009) A parametric architecture for tags clustering in folksonomic search engines, in Intelligent Systems Design and Applications, ISDA'09, Ninth International Conference on, pp. 279–282, IEEE
30. Ding Y, Jacob EK, Fried M, Toma I, Yan E, Foo S, Milojevi S (2010) Upper tag ontology for integrating social tagging data. J Am Soc Inf Sci Technol 61(3):505–521

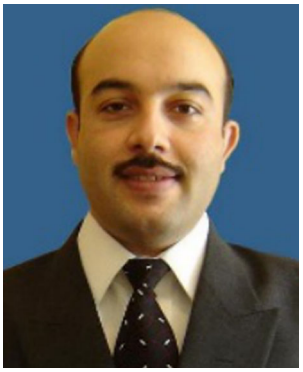
31. Ebrahimi T (2014) Security and Trust in social media networks. available : http://www.dmpf.org/obrainstorming/TouradjEbrahimiS.pdf
32. Echarte F, Astrain JJ, Córdoba A, Villadangos J (2004) Ontology of Folksonomy: A New Modeling Method, Conference'04,vol.289, ACM
33. Eda T, Yoshikawa M, Uchiyama T, Uchiyama T (2009) The effective- ness of latent semantic analysis for building up a bottom-up taxonomy from folksonomy tags. World Wide Web 12:421–440
34. Federica C, Antonina D, Pasquale L, Julita V (2013) Perspectives in semantic adaptive social web. ACM Trans Intell Syst Technol (TIST) 4(4)
35. flickr wrappr-precise photo association. [online] available: http://www4.wiwiss.fu-berlin.de/flickrwrappr/ (2011)
36. Fujimura K, Fujimura S, Matsubayashi T, Yamada T, Okuda H (2008) Topigraphy: visualization for large-scale tag clouds, in Proceeding of the 17th international conference on World Wide Web, WWW'08, pp. 1087–1088, ACM
37. Garcia A, Szomszor M, Alani H, Corcho O (2009) Preliminary results in tag disambiguation using dbpedia, in Knowledge Capture (K-Cap'09) - First International Workshop on Collective Knowledge Capturing and Representation - CKCaR'09
38. García-Silva A (2012) Discovering tag semantics. [online]. available: http://grafias.dia.fi.upm.es/Sem4Tags/
39. Geir S, Atle GJ (2011) Mining tag similarity in folksonomies, in proceedings of the 3rd international workshop on Search and mining user-generated contents, pp 53–60, ACM
40. Giannakidou E, Kompatsiaris I, Vakali A (2008) SEMSOC: SEMantic, SOcial and Content-based Clustering in Multimedia Collaborative Tagging Systems, in IEEE International Conference on Semantic Computing, pp .128-135
41. Giles J (2005) Internet encyclopaedias go head to head. Nature 438:900–901
42. Gobbo F :Improving flickr discovery through wikipedias.(2008)
43. Golbeck J, Koepfler J, Emmerling B (2011) An Experimental Study of Social Tagging Behaviour and Image Content. J Am Soc Inf Sci Technol 62(9):1750–1760
44. Gupta M, Li R, Yin Z, Han J (2010) Survey on social tagging techniques. SIGKDD Explor Newsl 12:58–72
45. Halpin H, Shepard H () Evolving ontologies from folksonomies: Tagging as a complex system. available http://www.ibiblio.org/hhalpin/homepage/notes/taggingcss.html/2012
46. Han Z, Mo Q, Liu Y, Zuo M (2010) Constructing taxonomy by hierarchical clustering in online social bookmarking, in Educational and Information Technology (ICEIT),2010 International Conference on, pp. V3–47 – V3–51,IEEE
47. Haridas M, Caragea D (2009) Exploring wikipedia and dmoz as knowledge bases for engineering a user interests hierarchy for social network applications, in On the Move to Meaningful Internet Systems, OTM 2009 R. Meersman, T. Dillon, and P. Herrero, (eds), vol. 5871 of Lecture Notes in Computer Science, pp. 1238–1245, Springer Berlin, Heidelberg
48. Harvey M, Baillie M, Ruthven I, Carman M (2010) Tripartite Hidden Topic Models for Personalised Tag Suggestion, Advances in Information Retrieval, 32nd European Conference on IR Research. ECIR 2010: 432–443
49. Hayati P, Potdar V (2009) Toward spam 2.0: An evaluation of web 2.0 anti-spam methods, in Industrial Informatics, INDIN 2009, 7th IEEE International Conference on, pp. 875–880, IEEE
50. Iijima C, Kimura M, Yamaguchi T: Implementing an image search system with integrating social tags and dbpedia, in Knowledge-Based and Intelligent Information and Engineering Systems R. Setchi, I. Jordanov, R. Howlett, and L. Jain (eds), vol. 6278 of Lecture Notes in Computer Science, pp. 264–272, Springer Berlin, Heidelberg.(2010)
51. Java A, Joshi A, Finin T (2008) Detecting communities via simultaneous clustering of graphs and folksonomies, in Proceedings of the Tenth Workshop on Web Mining and Web Usage Analysis (WebKDD), ACM
52. Javanmardi S, Ganjisaffar, Y Lopes CV, Baldi P (2009) User contribution and trust in wikipedia. Networking, Applications and Worksharing, CollaborateCom 2009, 5th International Conference on
53. Jung J.J : Matching multilingual tags based on community of lingual practice from multiple folksonomy: A preliminary result, in Trends in Applied Intelligent Systems Garca-Pedrajas,N., Herrera,F., Fyfe,C., Bentez,J., Ali,M. (eds), vol. 6097 of Lecture Notes in Computer Science, pp. 39–46, Springer Berlin, Heidelberg.(2010)
54. Kawakubo,H., Akima,Y., &Yanai,K: Automatic construction of a folksonomy-based visual ontology, Multimedia, International Symposium on, vol. 0, pp. 330–335. (2010)
55. KGVR Shankar R, Pudi V (2010) Frequent itemset based hierarchical document clustering using wikipedia as external knowledge, in Knowledge-Based and Intelligent Information and Engineering Systems R. Setchi, I. Jordanov, R. Howlett, and L. Jain, (eds), vol. 6277 of Lecture Notes in Computer Science, pp. 11–20, Springer Berlin / Heidelberg

56. Kittur A, Kraut R.E (2008) Harnessing the wisdom of crowds in wikipedia: quality through coordination, in Proceedings of the ACM 2008 conference on Computer supported cooperative work, CSCW'08, pp. 37–46, ACM
57. Ko¨rner C, Benz D, Hotho A, Strohmaier M, Gerd S (2010) Stop thinking, start tagging: tag semantics emerge from collaborative verbosity, in Proceedings of the 19th international conference on World wide web, WWW'10, pp. 521–530, ACM
58. Kobilarov G, Bizer C, Auer S, Lehmann J (2009) Dbpedia - a linked data hub and data source for web applications and enterprises, in proceedings of developers Track of 18th International World Wide Web
59. Körner C, Kern R, Grahsl HP, Strohmaier M (2010) Of categorizers and describers: An evaluation of quantitative measures for tagging motivation, in Proceedings of the 21st ACM conference on Hypertext and hypermedia, pp. 157–166, ACM
60. Koutrika G, Effendi FA, Gyo¨ngyi Z, Heymann P, Garcia- Molina H (2008) Combating spam in tagging systems: An evaluation. ACM Trans Web 2(4):1–34
61. Krause B, Schmitz C, Hotho A, Stumme G: The anti-social tagger: detecting spam in social bookmarking systems, in Proceedings of the 4th international workshop on Adversarial information retrieval on the web, AIRWeb'08, pp. 61–68, ACM.(2008)
62. Lee K, Kim H, Jang C, Kim HJ (2008) Folksoviz: a subsumption- based folksonomy visualization using wikipedia texts, in Proceeding of the 17th international conference on World Wide Web, WWW'08, pp. 1093–1094, ACM
63. Lee K, Kim H, Shin H, Kim HJ (2009) Folksoviz: A semantic relation-based folksonomy visualization using the wikipedia corpus, Software Engineering, Artificial Intelligence, Networking, and Paral- lel/ Distributed Computing. ACIS Int Conf on 0:24–29
64. Lee K, Kim K, Shin H, Kim H (2009) Tag sense disambiguation for clarifying the vocabulary of social tags, in Computational Science and Engineering,CSE'09, International Conference on, vol. 4, pp. 729–734, IEEE
65. Lee S, Neve WD, Ro YM (2010) Tag refinement in an image folksonomy using visual similar and tag co-occurrence statistics. J Signal Process:Image Commun 25:761–773
66. Lee SS, Yong HS (2007) Tagplus: A retrieval system using synonym tag in folksonomy, in Multimedia and Ubiquitous Engineering, MUE'07, International Conference on, pp. 294–298
67. Lifshits Y (2007) Web mining: Blogspace and folksonomies, A Guide to Web Research: Lecture 3
68. Lin H, Davis J: Computational and crowd sourcing methods for extracting ontological structure from folksonomy, in The Semantic Web: Research and Applications Aroyo,L., Antoniou,G., Hyvnen,E., ten Teije,A., Stuckenschmidt,H., Cabral,L., Tudorache,T (eds), vol. 6089 of Lecture Notes in Computer Science, pp. 472–477, Springer Berlin, Heidelberg.(2010)
69. Liu B, Zhai E, Sun H, Chen Y, Chen Z (2009) Filtering spam in social tagging system with dynamic behavior analysis, in Social Network Analysis and Mining, ASONAM'09, International Conference on Advances in, pp.95-100, IEEE
70. Loong J (2012) Folksonomy, tag collison and tag spam. Available : http://www.networksolutions.com/blog/2009/03/folksonomy-tag-collisions-tag-spam/
71. Lops P, de Gemmis M, Semeraro G, Musto C, Narducci F (2013) Content-based and collaborative techniques for tag recommendation: an empirical evaluation. J Intell Inf Syst 40:41–61
72. Lu C, Chen X, Park EK (2009) Exploit the tripartite network of social tagging for web clustering," in Proceeding of the 18th ACM conference on Information and knowledge management, CIKM'09, pp. 1545–1548, ACM
73. Maguitman AG (2005) Algorithmic detection of semantic similarity, in Proceedings of the 14th international conference on World Wide Web
74. Marchetti A, Tesconi M, Ronzano, F Rosella M, Minutoli S (2007) Semkey: A semantic collaborative tagging system, in Workshop on Tagging and Metadata for Social Information Organization at WWW, vol. 7, pp. 8–12
75. Markines B, Cattuto C, Menczer F: Social spam detection, in Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, AIRWeb'09, pp. 41–48, ACM.(2009)
76. Markines B, Cattuto C, Menczer F, Benz D, Hotho A, Gerd S (2009) Evaluating similarity measures for emergent semantics of social tagging, in Proceedings of the 18th international conference on World wide web, WWW'09, pp. 641–650, ACM
77. Mathes A (2004) Folksonomies - cooperative classification and communication through shared metadata. Comput Mediated Commun 47(10):1–13
78. Mika P (2007) Ontologies are us: A unified model of social networks and semantics, Web Semantics: Science. Serv Agents World Wide Web 5(1):5–15
79. Milicic V (2008) W3c semantic web use cases and case studies case study: Faviki
80. Min QX, Uddin MN, Jo GS (2010) The wordnet based semantic relationship between tags in folksonomies, in Computer and Automation Engineering (ICCAE). Int Conf IEEE 2:815–819

81. Mirizzi R, Ragone A, Di Noia T, Di Sciascio E: Ranking the linked data: The case of dbpedia, in Web Engineering B. Benatallah, F. Casati, G. Kappel, and G. Rossi (eds), vol. 6189 of Lecture Notes in Computer Science, pp. 337–354, Springer Berlin, Heidelberg.(2010)
82. Mirizzi R, Di Noia T: From exploratory search to web search and back, in Proceedings of the 3rd workshop on Ph.D. students in information and knowledge management, PIKM'10, pp. 39–46, ACM.(2010)
83. Mirizzi R, Ragone A, Di Noia T, Di Sciascio E: Semantic wonder cloud: Exploratory search in dbpedia, in Current Trends in Web Engineering F. Daniel and F. Facca (eds), vol. 6385 of Lecture Notes in Computer Science, pp. 138–149, Springer Berlin, Heidelberg.(2010)
84. Mirizzi R, Ragone A, Noia T, Sciascio E (2010) Semantic tag cloud generation via dbpedia, in E-Commerce and Web Technologies (Aalst,W, Mylopoulos, J, Sadeh, NM, Shaw, MJ, Szyperski, C, Buccafurri, F, Semeraro, G eds.), vol. 61 of Lecture Notes in Business Information Processing, pp. 36–48, Springer Berlin Heidelberg
85. Mo Q, Han Z, Liu Y, Duan D (2010) Analyzing tags in online social bookmarking systemes, in Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on, pp. V2–164 – V2–168
86. Moldvay J, Bax I, Frerichs A, Schuh M (2010) Tagmantic: A social recommender service based on semantic tag graphs and tag clusters, in Proceedings of the fourth ACM conference on Recommender systems, RecSys'10, pp. 345–346, ACM
87. Mousselly-Sergieh H, Döller M, Egyed-Zsigmond E, Gianini G, Kosch H, Pinon J (2014) Tag Relatedness Using Laplacian Score Feature Selection and Adapted Jensen-Shannon Divergence. MultiMed Model Lect Notes Comput Sci 8325:159–171
88. O'Brien JM (2006) The race to create a'smart' google
89. Pan R, Dolog P, Xu G (2013) KNN-Based Clustering for Improving Social Recommender Systems, in Agents and Data Mining Interaction, ADMI 2012, LNAI 7607. Springer, Berlin Heidelberg, pp 115–125
90. Passant A (2007) Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs, in Proceedings of the First International Conference on Weblogs and Social Media (ICWSM). Boulder, Colorado
91. Pirrone R, Piptone A, Russo G (2010) Semantic sense extraction from wikipedia pages, Human System Interactions (HSI), 2010 3rd Conference on, pp. 543–547
92. Poorgholami M, Jalali M, Rahati S, Asghari T (2013) Spam detection in social bookmarking websites, in Software Engineering and Service Science (ICSESS), 4th IEEE International Conference on, pp. 56–59, IEEE
93. Quattrone G, Capra L, Meo, PD Ferrara E (2011) Effective Retrieval of Resources in Folksonomies Using a New Tag Similarity Measure, in proceedings of the 20th ACM international conference on Information and knowledge management, pp. 545–550
94. Quattrone G, Ferrara E, De Meo P, Capra L (2011) Measuring Similarity in Large-scale Folksonomies,in proceedings (conf/seke/QuattroneFMC11), pp. 385–391
95. Rástočný K, Tvarožek M, Bielikova M (2013) Web Search Results Exploration via Cluster-Based Views and Zoom-Based Navigation, Journal of Universal Computer Science, vol. 19, no. 15
96. Resnik P (1995) Using information content to evaluate semantic similarity in taxonomy, in proceeding of 14th International Joint Conference on Artificial Intelligence, pp. 448–453
97. Ronzano F, Marchetti A, Tesconi M (2008) Tagpedia: a semantic reference to describe and search for web resources, in SWKM
98. Scheau C, Rebedea T, Chiru C Trausan-Matu,S: Improving the relevance of search engine results by using semantic information from wikipedia, in Roedunet International Conference, (RoEduNet), pp. 151–156.(2010)
99. Si X, Liu Z, Sun M (2010) Explore the Structure of Social Tags by Subsumption Relations, in Proceedings of the 23rd International Conference on Computational Linguistics, (coling 2010), Association for Computational Linguistics, pp. 1011–1019
100. Sigurbjörnsson B, Van Zwol R (2008) Flickr tag recommendation based on collective knowledge, in proceeding of the 17th international conference on World Wide Web - WWW '08, pp-327-336
101. Simpson E Clustering tags in enterprise and web folksonomies, Association for the Advancement of Artificial Intelligence, available: http://www.aaai.org/2012
102. Solskinnsbakk G, Gulla JA (2010) A hybrid approach to constructing tag hierarchies. OTM 2010, Part II, LNCS 6427, pp. 975–982, Springer-Verlag Berlin Heidelberg
103. Stampouli A, Giannakidou E, Vakali A: Tag disambiguation through flickr and wikipedia, in Database Systems for Advanced Applications Yoshikawa,M., Meng,X., Yumoto,T., Ma,Q., Sun,L., Watanabe,c (eds), vol. 6193 of Lecture Notes in Computer Science, pp. 252–263, Springer Berlin, Heidelberg.(2010)
104. Strohmaier M, Körner C, Kern R (2012) Understanding why users tag: A survey of tagging motivation literature and results from an empirical study. Web Semant Sci Serv Agents World Wide Web 17:1–11
105. Strube M, Ponzetto SP (2006) Wikirelate! computing semantic relatedness using wikipedia, in proceedings of the 21st national conference on Artificial intelligence, vol. 2, pp. 1419–1424, AAAI Press
106. Sung K, Kim S.C, Kim S.K: Tag Quantification for Spam Detection in Social Bookmarking System, in Advanced Information Management and Service (IMS),6th International Conference on, pp.297-303, IEEE.(2010)

107. Tan S.S, Kong T.E, Sodhy G.C: Annotating wikipedia articles with semantic tags for structured retrieval, in CIKM-SWSM, pp. 17–24.(2009)
108. Tang J, Leung H, Luo Q, Chen D, Gong J (2009) Towards Ontology learning from folksonomies. IJCAI 9: 2089–2094
109. Tibely G, Pollner P, Vicsek T, Palla G (2012) Ontologies and tag-statistics, New Journal of Physics,vol, 14, no.5
110. Tomuro N, Shepitsen A (2009) Construction of disambiguated folksonomy ontologies using wikipedia, in Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, People's Web'09, (Stroudsburg, PA, USA), pp. 42–50, Association for Computational Linguistics
111. Trabelsi C, Ben Jrad A, Ben Yahia S (2010) Bridging folksonomies and domain ontologies: Getting out non-taxonomic relations, in Data Mining Workshops (ICDMW), 2010 I.E. International Conference on, pp. 369–379
112. Trant J (2009) Studying social tagging and folksonomy: a review and framework. J Digit Inf 10(1)
113. Tsui E, Wang WM, Cheung CF, Lau AS (2007) A concept–relationship acquisition and inference approach for hierarchical taxonomy construction from tags. Inf Process Manag 46(1):44–57
114. Uddin MN, Duong TH, Nguyen NT, Qi XM, Jo GS (2013) Semantic similarity measures for enhancing information retrieval in folksonomies. Expert Syst Appl 40(5):1645–1653
115. Wu C, Zhou B (2009) Semantic relatedness in folksonomy, in New Trends in Information and Service Science, NISS'09, International Conference on, vol. 0, pp. 760–765
116. Wu C, ZHOU B (2011) Tags are related: Measurement of semantic relatedness based on folksonomy network. Comput Inform 30:165–188
117. Xia Z, Peng J, Feng X, Fan J (2014) Automatic Abstract Tag Detection for Social Image Tag Refinement and Enrichment. J Signal Process Syst 74(1):5–18
118. Xu Z, Liu Y, Mei L, Hu C, Chen L (2014) Generating temporal semantic context of concepts using web search engines. J Netw Comput Appl 43:42–55
119. Xu H, Wang J, Hua X, Li S (2009) Tag Refinement by Regularized LDA. Proceeding of 17th ACM International Conference on Multimedia, pp. 573–576
120. Xu G, Zong Y, Jin P, Pan R, Wu Z (2013) KIPTC: a kernel information propagation tag clustering algorithm, Journal of Intelligent Information Systems, Springer
121. Yago-project. [online] . available: http://en.citizendium.org/ (2011)
122. Yang H.C, Lee C.H: Automatic Detection of Social Tag Spams Using a Text Mining Approach, In Advances in Social Networks Analysis and Mining (ASONAM), International Conference on, pp. 441–445, IEEE.(2010)
123. Yao J, Cui B, Huang Y, Zhou Y (2010) Detecting bursty events in collaborative tagging systems, in ICDE, pp. 780–783
124. Yeung CM, Gibbins N, Shadbolt N (2007) Mutual contextualization in tripartite graphs of folksonomies. In: Aberer K, Choi K-S, Noy N, Allemang D, Lee K-I, Nixon L, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber G, Cudr-Mauroux P (eds) The Semantic Web, vol 4825, Lecture Notes in Computer Science. Springer, Berlin, pp 966–970
125. Yeung CA, Gibbins N, Shadbolt N (2008) Collective user behaviour and tag contextualisation in folksonomies, in Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 03, pp. 659–662, IEEE
126. Yoo D, Choi K, Suh Y, Kim G (2013) Building and evaluating a collaboratively built structured folksonomy. J Inf Sci 39(5):593–60
127. Yoo D, Suh Y (2010) User-categorized tags to build a structured folksonomy. Commun Software Netw, Int Conf 0:160–164
128. zemanta-a revolutionary new platform for accelerating online content production for any web user. [online] available: http://www.zemanta.com/ (2011)
129. Zhai E, Ding L, Qing S: Towards a Reliable Spam-Proof Tagging System. IEEE Fifth International Conference on Secure Software Integration and Reliability Improvement, pp. 174–181.(2011)
130. Zhai E, Sun H, Qing S, Chen Z (2009) Spamclean: Towards spam- free tagging systems, Computational Science and Engineering. IEEE Int Conf 4:429–435
131. Zhang H, Korayem M,You E, Crandall, DJ (2012) Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities, in proceedings of WSDM'2012, ACM
132. Zhang H, Korayem M, You E, Crandall DJ (2012) Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities, in Proceedings of the fifth ACM international conference on Web search and data mining, pp. 33–42, ACM
133. Zhou N, Cheung WK, Qiu G, Xue X (2011) A hybrid probabilistic model for unified collaborative and content-based image tagging. IEEE Trans Pattern Anal Mach Intell 33(7):1281–1294

Fouzia Jabeen is Ph.D. Scholar at the Department of Computer Science, University of Peshawar, Pakistan. Her research interests include Web Semantics, Information Retrieval, Folksonomies and Social Tagging Applications.



Dr. Shah Khusro received his Ph.D. degree from Vienna University of Technology, Vienna, Austria in 2007. He is currently working as Assistant Professor at the Department of Computer Science, University of Peshawar, Pakistan. His research interests include Web Semantics, Web Engineering, Information Retrieval, Augmented Reality, Ambient Assisted Living and Mobile Technology for People with Special Needs. He is currently working on several research projects in these areas. He has several publications in international journals and conferences.

Amna Majid obtained her Bachelor's degree from the Department of Computer Science, University of Peshawar, Pakistan. Currently she is pursuing her M.S. in Web Semantics from the same institute. Her research interests are Web Semantics, Folksonomies and Information Retrieval.

Dr. Azhar Rauf received his Ph.D. degree from Colorado Tech University, Colorado, USA. He is currently working as Assistant Professor at the Department of Computer Science, University of Peshawar, Pakistan. He has also worked as an IT consultant for Great West Healthcare, Bearing Point and Oracle Cooperation in United States of America. His research interests include Data Warehousing, Database Security and Big Data Analysis.