CHAPTER 2

OVERVIEW OF TAG RECOMMENDATION

2.1 BLOGS

Blog is considered as a part of a website. Typically, blogs are managed with regular updates by individuals. It includes videos, audio, music, graphics, commentaries, text, images, etc. Blogs are more interactive and it can be considered as a verb that adds content to a blog. Blogs can be run using blog software, or on usual web hosting services (Gemmel et al 2012). Process of social bookmarking takes place via tags which is shared by user for online resources. Shared tags provide useful metadata for retrieval systems. Collections of blogs are called blogosphere. Blogs are interrelated publicly via blogrolls, annotations, linkbacks and backlinks. Blog search engines are used mainly to search the contents of the blog.

There are many blog search engines available such as bloglines, BlogScope and Technorati. Blog posts can be in a type of feedbacks, opinion, reviews, comments, videos, etc. It includes a link to the preferred site with keywords. Technorati is a popular blog search engine which renders up-to date information on tags and searches that are very useful in categorizing the blog posts. It discovers the creator tagged blog records. Social bookmarking is a process which helps users find, build up, classify and manage information from the network. Bookmarks are common to all that can be shared with others. An user can save bookmark link information to get it easily. This will provide an important metadata for retrieval schemes. Tags
weight and font size denote its relevant frequency with other tags (Park et al 2011).

2.1.1 Types of Blogs

In social network environment, various types of blogs are available. But these types of blogs differ in their content types and way of written format (Herring et al 2004). The different types of blogs are explained as follows:

- Vlog: A blog consists of videos.
- Linklog: A blog with a links.
- Sketchblog: A blog contains collection of sketches.
- Tumblelogs: Blog includes little posts and various media categories.
- Photoblog: A blog comprising photos.
- Typecast blogs: Blogs which are written on typewriters and then scanned.
- Phlog: An infrequent blog is hosted on Gopher Protocol.
- Moblog: A blog created by a mobile phone.

2.2 BLOGOSPHERE

Social networking sites are described as a graph in blogosphere in which each node indicates the blogger or blog post. These social networking sites are mainly developed to know about others and to have touch with friends in society. Blogosphere allows blogger to create, annotate and share their information, thoughts and opinion or sentiment with others in a community. This provides the community experience to Blogosphere than the
other social networking sites. Anyone can create and maintain the blog to increase the blog popularity. Blogosphere provides the relevant information on how users create the blog contents, interests, expression and how to communicate with others (Chi et al 2007). Finding out the social interests of people is essential to connect users with common interests and support people to share their contents. On creating the reference structure in blogosphere, the blogs have been inter-related.

Blogosphere acts as graph in which each node is represented as blogs and the links represent as relations between blogs (Brooks and Montanez 2006). As a result, this graph can be changed according to time in improvement the internal and external relationships. Bloggers can share their personal opinions and render opportunities for improving business including marketing and enhance customer relationships. Blogosphere consists of richer information than blogs. Blog URL, Zhuang et al (2008) defines the blog profile and its chronological activities are recorded in series of blog state tuples. When blog post appears, blog state tuples are formed and added into their related blog profile with respect to their current time. If any post is not appeared, the blog profile is declined.

2.3 PURPOSE OF TAGGING

The main purpose of tagging is to categorize the web resources based on their content. If many users use the same word to tag an item, the tag will become large and bold (Strohmaier 2008). Tagging sites are constructed on the data which are produced by users designed for individual management to present other services like resource discovery. Tagging process is used for handling resources to an individual user. Tag recommendation supports a user to post his/ her blog by recommending latent related tags. Recommendation process is a greatest investigated scenario in Folksonomy context.


2.4 TAGGING SYSTEMS

Tagging has become a major infrastructure on the Web. Tagging is a process in which an user can give meaningful terms to a resource in online settings. An user allocates the keyword to a resource and it is called tagging. It allows users to annotate the particular web resources such as a web page, a blog post and a multimedia image with liberally selected keywords called tags. Tags are the topics for a resource and it is a non hierarchical keyword (i.e.) bookmarking, picture or file (Strohmaier et al 2010). Tagging is a classification method for blog objects and the act of matching related user defined keyword to a document. Therefore, tags are used to manage and categorize the content of web and it is very helpful in particular for searching content. In web 2.0 applications, large number of tagging systems is available including Delicious, Flickr, Bibsonomy, Technorati Last fm etc., Tagging system is evolved mainly for two purposes such as

- To systematize and manage the web content
- To perform searching, in order to discover the relevant content shared by other users.

2.5 TYPES OF TAGGING SYSTEMS

Depending upon the process of assigning tags to the web resources, tagging systems are broadly classified into two types such as simple tagging and collaborative tagging.

- Simple Tagging

In simple tagging, an user can only create tags for his own resources. Users in the network cannot create tags for the resources uploaded by the other users in the network. This type of tags is mainly used to improve
search and retrieval. For example, photos on Flickr.com, news on Digg.com or videos on Youtube.com.

- **Collaborative tagging**

  In collaborative tagging, several users can create tags for the same resources available on the network. It enables the users to assign tags to the blogs depending upon the content of the post. Therefore, the collection of tags assigned by the different users for single resources is called folksonomy. For example, Delicious, LibraryThing.com and CiteULike.org, LibraryThing.com allows collaborative annotations where resources (blogs, books, URL) can be annotated by the users who are interested to tag the resources (Golder and Huberman 2007).

### 2.6 AN APPROACH TO FOLKSONOMY

Collaborative tagging system is a social record storage area, where users have complete freedom to choose their tags and it also allows regular interaction with the entries posted by the other bloggers (Golder and Huberman 2007). Therefore, it leads to a situation of even single resource can possess large number of tags assigned by the different users (Lipczak and Milios 2010b). In order to organize the collection of assigned tags in predetermined structure, proper pseudo taxonomy called folksonomy is emerged. Folksonomy is defined as the pseudo hierarchy of tags where users created tags for the resources are managed through knowledgeable representation. Folksonomy is a newly coined phrase from the three words such as folks stands for users, taxis stands for classification and nomos stands for management (Solskinnsbak and Gulla 2009).

It is formally represented in the form of tripartite graph of hyper-edges that mainly relates three triples , disjoint sets such as i) set of users $u \in$
U , ii) set of resources \( r \in R \) and set of tags \( t \in T \). Folksonomy is mathematically expressed as \( F \subseteq U \times T \times R \). The main interpretation between these triples \{tag, user, URL\} is that user can create tag to the resources identified by URL (Zubiaga 2011).

2.7 REAL TIME TAGGING SYSTEMS

Depending upon the architecture and content type of web resources, tagging systems are classified into various numerous categories (Marlow et al 2006). In this approach, a small description about currently used representative tagging systems has been provided as follows:

- **Delicious:** It is a social bookmarking site that allows user to store and create the tag for web pages and resources.

- **Yahoo! MyWeb2.0:** It is also a social bookmarking system similar to Delicious, but web resources are shared among the users in social network of contacts.

- **CiteULike:** This tagging system allows users to create tags for the academic journals and research publications using the co-citations, Bibliography and references.

- **Flickr:** This tagging system is a photo sharing system that permits users to upload their own photos and to create tags for it. It also allows users to tag the photos of others shared in the network as well as maintains a network of contacts.

- **YouTube:** It is a video sharing system that agrees users to upload video content and also it can annotate uploaded video with meaningful tags.
- **ESP Game**: It is also important tagging system that makes internet game available online when users are randomly paired with each other. Users try to reduce the tags suggested by the other users when presented with a random photo.

- **Last.fm**: It is a music information sharing database that allows users to tag artists, albums, and songs.

- **Yahoo! Podcasts**: It is a regularly updated audio content tagging system that allows users to tag the audio content based on specific keywords.

- **Odeo**: It is another podcast information system same as that of Yahoo! Podcast but it supports searching along with tagging.

- **Technorati**: It is the most familiar tagging system for the blogs and also called weblog aggregator. Users can assign tags according to the content of the blogs posted by the users. Tags are used to perform searching in order to find the required blog with specific topic of interest.

- **Live Journal**: It is same as that Technorati tagging systems where users can annotate the tags for their own personal profile, posts and also create tags for the weblogs for other users in the community.

- **Upcoming**: It is a collaborative events database system where users can upload the information about the events that are supposed to be happened in future. And it also allows users to annotate the tags according to the content.
2.8 MOTIVATION FOR TAGGING

Tagging is a public and social activity, but users assign tags only to their own resources but not to the resources published by the other users in the networks. The main and primary intention is to motivate the user to assign the tags to web resources through providing proper incentives. The user is defined as the one who pursues the tagging and also can select the tag according to their thoughts, attitude and precise qualities. The motivation to tags is broadly classified into two types such as organizational and social practices (Gupta et al 2010). In Organization practice, tagging is in the form of structured filing where users are motivated to assign tags to develop a personal standard and also use the tags created by the other users in the network. Later in social practice, tagging is the form of communicative nature where users can create tags based on the certain quality of resources according to their own opinions. The following is a list of tagging incentives which suggests the probable motivations such as: future retrieval, contribution and sharing, attract attention, self presentation and opinion expression.

- Contribution and Sharing

  Tagging mainly motivates this feature through describing the resources efficiently. The conceptually related resources are appended into the clusters and unnecessary categories that are posted by either known or unknown audience are completely refined.

- Attract Interestingness

  Popular tags are mainly used to annotate their own resources because it can only attract the interest of other users in the network. Therefore, it may attract the interests of other users into their own resources.
If any information is available (tag clouds) to reflect the popularity of the tags, users are motivated to assign tags to the resources but it may create spam tags.

- Future Retrieval

The main aim of tagging is to make the future retrieval of resources easy. It is used to stimulate the activity of users through assigning the reminders to perform specific task. For instance, read tag can insist the activity of other users and this type of descriptive tags can provide the metadata information about the objects which is not explained in any other tags.

- Self Presentation

Self referential tags can supports this feature in tagging. It is used to create a unique identity of user among the available large number of resources. Also, each can make separate mark to represent their presence on particular resource. For example: “mystuff” can express the user particular identity on the web resources and then maintain the personal relationship with the corresponding resources.

- Opinion Expression

Tags also express their own opinion and also convey their value judgments for the resources shared by the other users in network. For instance, Yahoo’s Pod cast system provides the “elitist” tag which is mainly used to convey their opinion. In most of the occasions, people use tags to achieve the status in the large group of community.
• **Task Organization**

The important and appreciable feature of tagging is task organization. It is very easy to maintain the tags from web resources. Proper machine learning tools are used to classify and organize the similar tags according to the content of the web resources [for instance, “tdl” (to do list)].

• **Play and competition**

In order to create tags, users are motivated based on the set of internal or external rules. User can create the tags according to the certain rules that engaged for the corresponding system. For example, Odeo users can create most particular correct tags based on the rules mentioned in the tag clouds (Gupta et al 2010).

2.9 **DIFFERENT TYPES OF TAGS**

There are different types of tags available. The following section describes about the definition of various tags.

• **Content-Based Tags**

These tags provide the concrete description of the resource using precise terms. Therefore, it is very useful to determine the available content of the resource very effectively. Example: Automobiles, Odyssey.

• **Context-Based Tags**

It can provide the definite context of the resources where it was created or maintained. Mainly these tags present the information related to resources such as location and time. Example: 2012, 10:19, Los Angeles.
• **Subjective Tags**

Subjective tags are used to express the opinion and feedback of users in their perspective. Therefore, it is used to evaluate the quality of tags based on object recommendation. The main intention of this type of tag is to motivate the self expression of users in more convenient way.

• **Attribute Tags**

Attribute Tags are tags which are derived from the intrinsic attributes of resources and not from the content of the resources directly. These tags provide the information about resources. Therefore, it is mainly used to identify the characteristics and quality of the resources.

• **Organizational Tags**

These tags are used in order to remind the user to perform specific task with their personal stuff. It is not used for global tag aggregation along with the tags of other users. It is basically time sensitive tags and hence, it varies dynamically according to the time. It also suggests the user to actively engage with the specific task and also makes the user to connect with specific task according to their interest on perceived subject.

• **Ownership Tags**

Ownership tags provide information about the owner who possesses the resources.

• **Purpose Tags**

Purpose tags provide the general information about the resources. They are not derived from the content of resources but, derived according to
the main purpose of the resources. Therefore, these tags provide the information related to the user which is mainly required by the user. (Example: Latex tags are the purpose tags that can provide recommendation for music and also to translate text).

- **Factual Tags**

  Factual tags are the tags used to identify the facts related to the objects such as user, location and concepts. These tags are mostly agreed by many users and also, it can be applied to any type of object. The aforementioned tags such as content based, objective, attribute and context based tags falls under the factual tags category. It is mainly used to obtain information about the resource in order to perform specific task.

- **Personal Tags**

  Personal tags are the tags where user only listener of the tags that was applied by themselves. It is mainly used to organize the objects of users such as (user ownership, task management and self identification).

- **Self-referential tags**

  Self referential tags are tags which refer to the resources of themselves. Large numbers of similar tags are posted but they are intended to tag only one specific resource.

- **Tag Bundles**

  Tag bundle is also a type of tags which was created on grouping the large number of tags in the form of hierarchical folksonomies. Several taggers
have selected tag URLs which is the fundamental web address for the server (Xu et al 2006).

2.10 ENHANCING TAGGING SYSTEMS

In social tagging, tags are free form labels assigned by any users available on the network and also it is not derived from any controlled vocabulary (Frank et al 2009). Therefore, in order to provide more meaningful terms as tags, tagging systems supports various features. It can predict the suspicious tags and correct them. The list of necessary features is represented as follows:

- Missed tags

  If users are supposed to miss the tags that are related to the content of the web resources, then tagging system suggests the additional tags extracted from the concept extraction technique.

- Spell check

  In order to mark the misspelled tags, open source Spell checker software (Aspell) is associated with tagging systems. It can easily mark the suspicious tags and also suggest the same tags with correct spelling.

- Unrelated Tags

  Standard machine learning algorithm such as concept extraction technique is deployed in tagging system. It can determine the tags that are unrelated to the content of the web resources tags and are separated.
• **Preferred form of tags**

If any tags require any preferred form, tagging system must provide suggestions for the forms such as capitalization, plural form of the tags to the user.

• **Literal meaning for tags**

In tagging system, a lexical dictionary tool called WordNet is deployed to suggest the synonyms of the tags.

• **Relation between tags**

Ontology learning techniques are supported by tagging system to track the more general concepts in the previously derived tags. It can determine the relation between the previously derived tags. Thus, more number of conceptually related tags is suggested from the basic level.

### 2.11 **LIMITATIONS IN COLLABORATIVE TAGGING SYSTEMS**

Despite the simplicity and popularity of collaborative tagging as an information organization approach, Tags tend to be noisy and sparse due to the uncontrolled vocabulary and annotation comes at the expense of several limitations (Halpin et al 2007). Among that, three limitations are explained as follows:

• User creates tags to the resources based on their personal opinions, their knowledge background and their preferences. Furthermore, users may be describing the same object based on their different granularity. It leads to noisy tag space and
thus, makes it difficult to find the resources tagged by other users.

- In some cases, users utilize polysemous words (a word that has many related senses) in order to tag the web resources. The lack of semantic distinction in tags can lead to inappropriate connections between items. Another problem is that different tags, which are either synonymous or have closely related meaning increase data redundancy, leading to reduced recall of information.

- Due to the uncontrolled vocabulary and annotation guidelines, user is tending to assign a very small number of tags to an object.

2.12 TAG RECOMMENDATION BASED ON SOCIAL COMMENT NETWORK

The aim of tag recommendation is to provide a set of relevant tags according to the user requirement. The characteristic representation of users mainly relies on their general interest and social association is referred to as social comment network. Tag recommendation is developed based on the context of the social comment network. Social comment network allows users to communicate with new participants and to maintain relationship with anyone in the network. Group of members with similar interest are united together and form a community. They even share their interest about specific topic. The most admired activity of this social network is that a user can upload photos and videos and can receive comment over it. This system has
two stages such as developing comment network and also computes the author’s prestige (Jiang et al 2010).

2.12.1 Develop Comment Network

In media social network site, many users can comment on the uploaded resource. The user relationships can be represented in a directed graph, in which the nodes will represent a user and the link represents the relationship.

2.12.2 Calculate User Prestige

The prestige of the user is computed according to its prominence on social network. It is used to evaluate the popularity of resources shared by the user. Depending upon the popularity of the resources, corresponding prestige is computed. The popularity of the resources is mainly determined from the number of links associated with the corresponding resources. Therefore, prestige of every user is computed according to the prominence and they are ranked according to their computed prestige value. Users with high prestige score can form more number of groups and also tags can be added to the photos.

2.13 ASSOCIATION BETWEEN SOCIAL AND SEMANTIC WEB

Tags are recommended based on the association of semantic web on the social web. In social web, a user plays an important role to upload the resources over the web. It refers to the increased user participation on the web. On the other hand, Semantic web is guided under the control of World Wide Web Consortium (W3C) and also extended from the existing web and not from a separate web. The main aim of semantic web is to offer a
cooperative framework that provides a well defined meaning for the information available on the web. Blogs (Social web) and Topic ontology (semantic web) are associated together to identify the tag which is recommended to the blog users. The need of topic ontology in tag recommendation process is to provide conceptually relevant recommended tags to the content of the blog. Therefore, Topic ontology for a specific domain is efficiently constructed using online web resources such as Wikipedia and WordNet based on the relevant keywords and its lexical relationships. Topic ontology takes keywords from the content of the blog as input, in order to recognize relevant tags which are recommended to the blog users. Topic ontology assists users to recommend the possible tags with high score.

2.14 TOPIC ONTOLOGY

Topic ontology is mainly used to categorize the web pages based on their content (Zhou et al 2006). It is defined as ontology with a set of topics that are interconnected using semantic relations. It is also depicted as the graph in which each node represents the specific topic and nodes are connected using semantic relationship. Finally, it results in a topic hierarchy. It comprises a group of relevant concepts related to the specific domain (Maguitman et al 2010). A hierarchical semantic relationship is maintained among the concepts in the topics. Topic ontology is mainly used to categorize the content of web resources. In the aspect of tag recommendation, topic ontology is used to determine the topics. These topics are used as tags for corresponding web page. Topic ontology is constructed by using the set of topics extracted from Wikipedia and semantic relation between set of topics is derived from the lexical relation of wordnet. Hierarchical Construction of topic ontology is shown in Figure 2.1.
WIKIPEDIA

Wikipedia is a multilingual online encyclopedia which contains articles and categories that are related to one precise concept. The Internet users widely use the Wikipedia. In wikipedia, each web page defines the set of concepts related to the specified domain and is more formally organized in the form of hierarchical fashion according to the general knowledge. As mentioned before, Wikipedia offers category structure but for constructing topic ontology, the categories, links, attributes and topics to provide more explicit information to the users need to be extended. Therefore, Wikipedia has been used to develop topic ontology efficiently. Wikipedia can perform
several tasks such as content and form modification and also it is scalable and semantic in performance.

Wikipedia presents the ways to organize its contents. The organized structure provides good chance to formalize the topic ontology without any complexities and inconsistencies. On providing the web pages according to the different word senses, disambiguation between similar words can be avoided. Searching in Wikipedia is a straightforward approach and hence, set of keywords associated with each web page can able to construct topic ontology. But, keywords itself are not enough to construct topic ontology because they are just an arrangement of set of relevant topics related to specific domain. Moreover, it does not render any semantic relation among the set of concepts (Krötzsch et al 2007).

2.16 WORDNET

WordNet is a lexical reference system consisting of set of related usual language terms that falls under same lexical concepts. Each word possesses one or more similar synonyms depending upon the word sense such as nouns, verbs, adjectives and adverbs. A collection of words that distributes one sense is known as synset and it is organized in the form of hierarchy manner (Snasel et al 2005). It is mainly used to establish the semantic association among the group of synonyms to suggest the morphological relations among words. In order to construct ontology, WordNet affords the concept hierarchy with Hypernym association using top down and bottom up approach. Using this approach, all terms in hypernyms that are used from top level is called top down approach and the hypernyms terms that are used from top level upto one beyond the synset term to change terms with topics is called bottom up approach.
Furthermore, WordNet is connected to an Inter-lingual index and this index is used to inter-relate the similar topics. WordNet differentiates between word senses and words. Synset is a collection of words that share one sense and it discovers semantic concept. Meaning of words is represented using synset in WordNet. WordNet and topic ontology are exploited to identify the semantics of tags. Classifier uses information in Wikipedia and WordNet to anchor Wikipedia articles. The set of related terms in synset are interlinked by conceptual -semantic and lexical relationship such as Hyponym/Hypernym (subclass/super class, represented as IsA relation) and Meronym/Holonym (part-whole relationships) which is used to facilitate the efficient construction of topic ontology. The WordNet has been employed to extract semantic relationship between set of categories extracted from wikipedia.

2.17 SPREADING ACTIVATION ALGORITHM

Spreading activation, introduced by Collins & Loftus was applied in psycho linguistics and semantic priming. Soon after, spreading activation is effectively used in various research areas in computer science particularly in the process of information retrieval. The spreading activation algorithm is an effective approach than other logic reasoning approaches because operations performed are represented in the form of pulse. Each pulse has three phases such as i) Pre-adjustment ii) Spreading iii) post adjustment. Among the three phases, Preadjustment and Post adjustment phases can able to manage the activation score of every node and the whole network. It monitors the activation decay level of the nodes in the network (Dix et al 2010). Therefore, it can preserve the node from previous pulses and avoids conflict. The most important phase is spreading phase, because the activation value is spreaded among the nodes in the network. Various ways are available to spread the
activation value throughout the network. The input level of the node is computed as follows:

\[ I_j = \sum O_j \times W_{ij} \]  

(2.1)

Where  
\( I_j \) = input for the node \( j \)  
\( O_j \) = Output of node \( j \) which is linked to the unit level of node \( i \).  
\( W_{ij} \) = weight associated between the two nodes \( i \) and \( j \).

The final activation score level of the node after reaching the destination is computed from the starting activation level and the number of links passed by the node before reaching the destination. The distance taken to reach the destination is needed to determine the activation score from the specified distance from the node. In order to reduce the spreading throughout the network during activation process, special attention is provided on those nodes which are connected to the large number of nodes in network. In the association information retrieval, activation process is carried out based on the specified inference rules (Crestani 1997).

2.18 APPLICATIONS OVERVIEW

While recommending tags to the user, resource spreading activation algorithm is geared to the applications such as spam reduction, sentiment analysis and tag popularity. Initially, this approach detects spam emails, spammers, messages, web spam etc for blog security according to the interest score which is applied on the tags. Irrelevant tags are reported as spam in spam reduction process. Secondly, this work analyses the area of sentiment analysis or opinion mining in which user may express their sentiments or opinions in diverse ways and attitude. Algorithm determines the user opinion expressed in a blog may be positive, negative or neutral. Finally, tag popularity is calculated based on the co-occurrences of tags in a blog. Interest
score is applied on each tag to find out how many times it occurred on blog posts. Based on the highly activated scores, tags are represented as most popular tags.

2.18.1 Spam

Tag popularity and high bookmarking page rank are misused by spammers. Spammers post their entries and send same post many a times from various accounts. Spammers try to enhance the ranking but not the page content quality. Thus, it increases the web sites traffic and interrupts the services. Keyword which is added by the spammers would not be matched with the bookmark content (Krause et al 2008).

A Spammer adds others as friends instead of giving friend request to decrease the threat of being caught as spammer since they get lesser friend request than genuine users. A Spammer may be a program and not a person (Jin et al 2011). Strategy attackers may attack the anti spam approach assigned by the system to detect the tag spam (Liu et al 2009). These attackers often save particular URLs related to tag spam in popular taxonomy of Delicious. This approach is useful to detect and avoid these kinds of strategic attackers and spam from the blogs and to provide an effective tag recommendation.

2.18.2 Sentiment Analysis

This approach identifies the bit of text stating a positive or negative opinion in sentiment analysis. Opinion is categorized into two contrasting sentiment divergence. Sentiment associated (i.e.) positive, negative and neutral categories are called polarity categorization. Categorizing a topic into good or bad is considered as opinion classification task (Pang and Lee 2008). Positive represents the top scored topics and negative represents the less
scored topics in terms of positive and negative sentiment (Melville et al 2009). On concentrating into the subjective sentences, this approach provides the increased sentiment analysis accuracy. Sentiment analysis expresses users’ feelings through tags. In this research work, widespread emotions are identified and provided as output based on the collections of tags. Each tag is assigned with interest score that convey the emotion’s weight. In order to identify the sentimental tag, the lexicon, WordNet that directly refers to the sentimental tags are incorporated (Baldoni et al 2012).