Overview

Data mining-based tag recommendation system: an overview

Subramaniyaswamy Vairavasundaram,[1]* Vijayakumar Varadharajan,[2] Indragandhi Vairavasundaram[3] and Logesh Ravi[1]

The advent of high-speed Internet connections has revolutionized the way research is being carried out to obtain relevant information. Conversely, retrieving pertinent information from the copious resources available is not only difficult but also time consuming. In the recent years, tagging activity has been perceived as a potential source of knowledge on personal preferences, interests, targets, goals, and other attributes. Tags allow users to effectively annotate resources using keywords to personalize their recommendations and organize the resources for easy retrieval. However, the preference of users varies extremely resulting in tagging being counterproductive. These shortcomings reduce the application of the tagging system for filtering as well as retrieval of information. The tag recommendation system becomes useful by suggesting a set of relevant keywords to annotate the resources. This paper presents a review of the tag recommendation systems and the constraints that affects the available tag recommendation systems. Furthermore, we propose the use of spreading activation algorithm to study the role of constructed topic ontology for efficient tag recommendations. This approach is founded on the assumption that tags that are recommended to the user are predicted from the extracted keywords from the existing blogs and the topics in constructed topic ontology. We have also proposed a tag classification system, namely Correlation-based Feature Selection–Hybrid Genetic Algorithm and classifier HGA-SVM (support vector machine), and have compared the results with results produced by other existing feature selection methods. The results obtained from the experiments have been presented.

How to cite this article:
WIREs Data Mining Knowl Discov 2015, 5:87–112. doi: 10.1002/widm.1149

INTRODUCTION

Social data mining has gained interest among researchers and practitioners in the recent past all over the world. Social Web or Web 2.0 has changed the way the end users use the Internet, e.g., instead of being consumers, users have now become contributors of knowledge.[1] Through social networking sites, users define their close networks with whom they intend to keep in touch. These sites enable the users to share content with each other, create discussions, and manage the networks by allowing access to a restricted community or everyone who visits the web. Sharing of content is not limited only to text, but a great many contents such as pictures, videos, slides, trips, and so on are being shared. However, sharing the content alone will not achieve the goal of reaching the target audience, but making it visible is crucial. For the content to be discoverable by other users, many of the websites allow the users to add

*Correspondence to: vsubramaniyaswamy@gmail.com

[1]School of Computing, SASTRA University, Thanjavur, India

[2]School of Computing Science and Engineering, VIT University, Chennai, India

[3]School of Electrical and Electronics Engineering, SASTRA University, Thanjavur, India

Conflict of interest: The authors have declared no conflicts of interest for this article.

tags or keywords to resources. Tags are thus used to navigate and find resources which browsers are searching for. Tagging systems can influence the way a search is made, increase reputation, improve personal organization, and detect spam.[2] At the same time, it can introduce opportunities for data mining and new methods of social communication. The web sites provide some tag recommendation mechanisms that make the process of finding good tags for a resource easy further to strengthening the tag vocabulary among users. For example, Del.icio.us had a tag recommender (http://www.socio-kybernetics.net/saurierduval/archive/2005_06_01_archive.html) and resource recommendation system (http://blog.del.icio.us/blog/2005/08/people_who_like.html) in place as early as 2005.[69] Nevertheless, algorithmic details of these resources were not published. Perhaps, the recommendations provided for those tags might have given suggestions based on the most frequently assigned tags to the resources. Despite being used widely on the social web, tagging system still has several unresolved issues due to its heterogeneity, ambiguity, and inadequate organization between tags. Furthermore, not many effective recommendation strategies exist in practice.

Therefore, in this study, we propose a process that can effectively recommend tags using topic ontology approach. The chosen topic ontology-based approach has been shown to be superior to the effective collaborative approach, as we recommend proper guidelines to generate efficient tag recommendations. This article reviews the available tag recommendation systems that have been used for extending the capabilities of tag recommender systems. Constraints of the recent generation of tag recommendation systems and potential extensions that can present enhanced recommendation capabilities are also considered. This study also addresses the technology used to produce recommendations with special focus on the use of data mining techniques.

This article is organized into seven main sections. *Tagging Systems* section briefly describes the tagging systems with special reference to the motivation behind tagging, Folksonomy, and its types. *Tag Generation Models* section analyses the different models of the tag generation and has compared and contrasted the models to establish a foundation for the necessity of this research. *Types of Tag Recommendation Systems* section deals with the review of tag recommendation systems. *Proposed Tag Recommendation Using Spreading Activation Algorithm* section provides the details about the method of tag recommendation system and discusses the results obtained. Proposed Tag Classification System section discusses the methodology used for proposed tag classification system and *Conclusion* section highlights the conclusion and recommendations for future work.

TAGGING SYSTEMS

Tagging is a process in which a user can give meaningful terms to a resource to facilitate the easy discoverability of the resource. Tags are the nonhierarchical keywords of a resource, i.e., bookmarking, picture, or file.[3] Tagging allows the user to categorize the web resources, such as web pages, blog posts, photos, multimedia images, and so on, based on their content. For example, if many users use the same word to tag an item, the tag will become more popular.[4] Tagging sites are constructed with the data that are produced by users designed for individual management and resource discovery. Thus, the main objective of the tagging system is to structure and manage the web content and to discover the relevant content shared by other users. In web 2.0 applications, a large number of tagging systems is available, e.g., Delicious, Flickr, BibSonomy, Technorati Last fm, and so forth.

Depending upon the process of assigning tags to the web resources, tagging systems are broadly classified into two types: simple tagging and collaborative tagging. In simple tagging, users create tags for resources created by them but others in the network will not be able to add tags to those resources. Simple tagging is mainly used to improve search and retrieval. Examples are photos on Flickr.com, news on Digg.com, and videos on Youtube.com.

In collaborative tagging, more than one user 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.[73] The collection of tags assigned by different users to a single resource is called Folksonomy. For example, among Delicious, LibraryThing.com, and CiteULike.org, LibraryThing.com allows collaborative annotations where resources (blogs, books, and URL) can be annotated by the users who are interested to tag the resources.[5]

Based on the architecture and content of web resources, tagging systems are classified into numerous categories.[2] Few of the representative tagging systems are Delicious, Yahoo! MyWeb2.0, CiteULike, Flickr, YouTube, ESP Game, Last.fm, Yahoo! Podcasts, Odeo, Technorati, Live Journal, and Upcoming.

Folksonomy

Collaborative tagging system is a social record storage area, where users have complete freedom to choose

their tags and allows regular interaction with the entries posted by other bloggers.[5] Thus, a single resource can possess large number of tags assigned by the different users.[6] The collection of assigned tags is organized in a predetermined structure using pseudo taxonomy called Folksonomy. Folksonomy is defined as the pseudo hierarchy of tags where user-created tags are managed through knowledgeable representation. It is a newly coined phrase, where folks stand for users, taxis for classification, and nomos for management.[7]

It is formally represented in the form of tripartite graph of hyper-edges that mainly relates three tuples, disjoint sets: (i) set of users $u \in U$, (ii) set of resources $r \in R$, and (iii) set of tags $t \in T$. Folksonomy is mathematically expressed as $F \subseteq U \times T \times R$, where U is a finite set of users; T, a finite set of tags; and R, a finite set of resources. Using the triples {tag, user, and URL}, user can create tag to resources identified by URL.[8]

Users are motivated to add tags to web sources, and they can either add tags as an organizational practice or as a social practice.[9] In organizational practice, tagging is done in the form of structured filing, where users are motivated to assign tags to develop a personal standard and also add tags to resources created by the other users in the network. In the social practice, tagging is the form of communicative nature, where users can create tags based on certain quality of resources according to their own opinions. An indepth study conducted by Ames and Naaman[10] and Marlow et al.[2] provide a comprehensive list of incentives that can be used to motivate the users to tag. Some of the incentives include future retrieval, contribution and sharing, attract attention, self-presentation, play and competition, task organization, and opinion expression.[9] Ames and Naaman[10] demonstrated that incentives given either in the form of rewards or entertainment, such as games, can motivate users to tag photos in mobile and online environments.

Categorization of Tags

Types of social tags are many. Therefore, an attempt has been made to differentiate the tags based on its relevance in information retrieval (IR) and knowledge management. These social tags, in addition to defining the content, represent subjective opinions,[11,12] self-presentation, contextual information,[10] and organization forms.[13] Earlier, the tag collections retrieved from social tagging systems such as Delicious, Flickr, and Last.fm were classified manually.[14] These classifications were necessary to study how the tags are distributed and used in improving the search. Cantador et al.[15] went a step further to categorize these tags automatically. They categorized the tags into four main types, namely, content-based, context-based, subjective, and organizational. Here, we categorize the tags based on the past studies.

Content-Based Tags

These tags provide concrete description of the resource using precise terms.[16] Thus, they contribute to the effective determination of the available content of the resource, e.g., Automobiles and Odyssey.

Context-Based Tags

Context-based tags provide the definite context of the resources, e.g., where they were created or maintained. Predominantly, these tags present the information related to resources, such as location and time,[9] e.g., 2012, 10:19, Los Angeles.

Subjective Tags

Subjective tags are used to express the opinion and feedback of users from their perspective. They are used to evaluate the quality of tags based on object recommendation.[13] The main intention of this type of tag is to motivate the self-expression of users in a more convenient way, e.g., awesome, wonderful, and handsome.

Attribute Tags

Attribute tags are those tags derived from the intrinsic attributes of resources and not from the content of the resources directly. These tags provide the information about resources.[16] Therefore, it is mainly used to identify the characteristics and quality of the resources, e.g., scary, stupid, and funny.

Organizational Tags

Organizational tags are used to remind the user to perform specific tasks with their personal stuff. They are not used for global tag aggregation along with the tags of other users.[9] These are basically time-sensitive tags; hence, they vary dynamically according to the time. They suggest the user to actively engage with the specific task and make the user to connect with specific task according to their interest on perceived subject, e.g., toread, goto, and jobsearch.

Ownership Tags

Ownership tags provide information about the owner who possesses the resources,[16] e.g., Amazon, Flickr, and Yahoo.

Purpose Tags

Purpose tags provide the general information about the resources. They are not derived from the content of resources, but from the main purpose of

the resources.[9] Therefore, these tags provide the information related to the user, which is specifically required by the user, e.g., 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 upon by many users and can be applied to any type of objects. The aforementioned tags, such as content-based, objective, attribute and context-based tags fall under the factual tags category. These are used to obtain information about the resource, in order to perform specific tasks,[13] e.g., people, concepts, and places.

Personal Tags

Personal tags are the tags where the user only listener of the tags that was applied by themselves. These are mainly used to organize objects of users, such as user ownership, task management, and self-identification,[16] e.g., mystuff, mybag, and mylaptop.

Self-Referential Tags

Self-referential tags refer to the resources themselves. Large numbers of similar tags are posted, but they are intended to tag only one specific resource,[9] e.g., sometaithurts.

Tag Bundles

Tag bundles are also a type of tag created by grouping a 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,[13] e.g., Tagging online courses to http://nptel.ac.in url.

TAG GENERATION MODELS

Tag generation models have evolved to analyze and understand the tags based on the content of the web resources. To generate tag recommendation, knowledge about the tagging system are essential, these models are created to obtain the background knowledge about the content of web resources.[17] Such models ensure that the current user is aware of the tags suggested by the previous users. Basically, selection of tags by the current user is influenced by the tags assigned by the previous user. Thus, already suggested tags become the main criteria to generate tags. Hence, much effort is not needed to generate tags. Information foraging is a new technique developed to adapt the suggested tags according to the user behavior.[18,78] Furthermore, it is also used to achieve faster retrieval of the information from web resources.

Polya Urn Generation Model

This model can express the variation of user constant prototype according to the time the predetermined collective tags are applied to the resources. Tags are generated from the previously suggested tags by any users in the network. This model does not support the generation of new tags. Adding generated tags into the systems is not feasible as it is supposed to be the same as that of already suggested tags. This stochastic model is described using the experiment where an urn consists of two balls with two different colors (e.g., a green and a blue ball). First, a ball of particular color is taken from the urn and dropped into another urn which has a ball of similar color. These steps are repeated for a number of times till the fraction of balls in each urn gets stabilized into random limits. However, for each execution of the experiment, this fraction of balls congregates to random, a limit that provides different predictable outcomes.[9] Polya urn with unbounded color numbers has many colors to address real life data, as we do not know the actual total number of "colors." So it is not advised to fix number of colors before observing the entire data. Even while observing the entire data population; there is no clue for color of the upcoming balls. This is more helpful in the other applications such as functional Magnetic Resonance Image (fMRI) pattern activation and research of brain image.

In general, Polya urn model is defined as,

$$G \sim DP\left(\alpha, G_0\right)$$
$$\text{color}\,(i) \sim G \forall i$$

where α is scalar that controls the one observing new color and G_0 is distribution over colors. This experiment can be allegorized to the tags that are generated to the web resources similar to the previously assigned one. Thus, this model cannot add the newly selected tags into the system.

Yule-Simon Generation Model

In order to overcome the problem in the Polya urn generation model, Yule-Simon generation model was proposed. In this approach, new tags are created for each step and combined with the tagging system with a low probability of p. Consequently, a large number of different tags are added to describe the web resources, which dynamically vary over time. But the number of tags generated at each step will be declined. Yule-Simon Generation Model deals with the data population of entities. Every entity has a property and this property is also known as elements, which is represented in the numeric type. In this model, single words are considered as entities and word usage count

is element of the particular text. Addition of entities for population generation is the main process in the Yule-Simon generation model. When the property modified entity is chosen along with probability proportional to the size of the property, results in property distribution. This model is described as follows: newly generated tags are added to the group of tags with probability r. Largely, the newly added tag is not supposed to occur anywhere in the system. The tags are generated using the probability p from the tags that are already requested. It can follow the frequency-rank distribution with a power law tail whose exponent is given by $a = 1 - r$. The number of different tags (N) generated into the system is directly proportional to the number of newly generated tags over time (T).[9] The generation rate of new tags decreases and, finally, reaches zero. This type of sub-linear growth is represented using Heap's law.

Yule-Simon Model with Long-Term Memory

Yule-Simon model is similar to that of Simon model, but applicable with long-term memory. In the Simon model, tags were generated using the copy of previously assigned tags with the same probability r. On the contrary, in this model, duplicate versions of previously assigned tags are generated using long-term memory. The model can be described as follows: users of a collaborative tagging system can assign tags to resources where text is constructed based on n number of words. The context of the user is efficiently described using the specified model that is used to identify the behavior of user. If new tags are generated by the system with probability p, after time step t, one word from already assigned tags are copied into the newly generated tags with probability $1 - p$. It can be performed for n number of steps. The probability decays follow the power law, which is expressed using a power law distribution function as $P(x) = n(t)/(x + (t))$, where $n(t)$ is a normalization factor, and t is the time-scale factor over probabilities. This approach can produce a characteristic value along the slope in the frequency-rank distribution of co-occurrence tag streams. However, this model fails to explain the distributive nature of resources among the tag streams and does not provide any information about the decay rate of the number of different tags.[19]

Information Value-Based Model

This tag generation model imitates previously assigned tags and it also chooses tags by analysing its information value too.[20] If the information value is 1, then the tag can be used for appropriate resources and when the information value is 0, then the tag cannot be used for any resources. Along with the selection of tags using information value, this model also imitates previously assigned tags using the Polya urn model. Preferential attachment models and linear combination of information value is used for the tag selection.

Probability of a tag y is represented as,

$$P(y) = \lambda \times P(I(y)) + (1 - \lambda) \times P(a) \times P(o) \times P\left(\frac{R(y)}{\sum R(i)}\right)$$

Here, in time (t) the probability of the user performing tagging action is $P(a)$ and with number of tags (n), $P(n)$ determines the number of tags distributed per action. Probability $P(o)$ is a constant used to represent the old tags. If old tag is used, probability $\frac{R(y)}{\sum R(i)}$ is added to the tag, where $R(y)$ is the number of times the tag was used in the past and $\sum R(i)$ is the aggregate sum of all the tags used previously. This model helps in a plain power law distribution of generated tags and its frequencies with a linear growth of the set of different tags. This model is not capable to reproduce the decaying growth of tags.

Enhancing Tagging Systems

In social tagging, tags are free from labels assigned by any users on the network, and they are not derived from any controlled vocabulary.[21] Tags should be meaningful to facilitate its prevalent use. Tagging systems offer certain features that could predict suspicious tags and correct them. A list of necessary features available in the tagging systems is given in the following sections.

Missed Tags

If users miss the tags that are associated with the content of the web resources, then tagging system suggests additional tags extracted through the concept extraction technique.

Spell Check

In order to mark the misspelled tags, open source Spell Checker software (Aspell) has been linked to the tagging systems. It can easily mark the suspicious tags and suggest the same tags with correct spelling.

Unrelated Tags

Standard machine-learning algorithm, such as concept extraction technique, is deployed in tagging system to determine the tags that are unrelated to the content of the web resources tags and separated.

Preferred Form of Tags
User preference is taken into consideration by suggesting capitalization, plural form of the tags, and so on to the user.

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

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

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 that appears at the expense of several limitations.[22] Of which, three limitations are explained as follows:

- Users create tag to the resources based on their preference, knowledge, background, and personal opinion. Furthermore, users may label the same object based on their different granularity too, thus resulting in noisy tag space, which makes it difficult to find the resources tagged by other users.
- In some cases, users utilize polysemous words (word having many related senses) to tag web resources. The absence of semantic difference in tags can lead to incorrect connections between items. Moreover, different tags, either synonymous or having closely related meaning, escalate data redundancy leading to less recall of information.

Due to the uncontrolled vocabulary and annotation guidelines, users tend to assign a very small number of tags to an object.

Purpose of Recommending Tags
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.[4] Tagging sites are constructed with the data that are produced by users designed for individual management to present other services such as 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.

The limitations in the appropriate usage of tags have lead researchers to develop techniques to automatically recommend the set of tags to the users. The purpose of recommending tags is many fold: (1) to automatically suggest popular and related tags for untagged resources to users directing them toward the information they request,[23] (2) to get accurate user preferences via tags and exploit it for efficient recommendation,[24] and (3) to act as mediator between users and resources, handling the user preferences in social networking sites or blogs.[74] By suggesting suitable tags, the tag recommendation system, improves the process of resource annotation.

In collaborative tagging systems, community services, such as Flickr, Delicious, and BibSonomy, are common types of Folksonomy. It is popular for content management and sharing of resources among millions of users. These community services allow users to add tags on online items, such as photos, videos, and text. Folksonomies allow a set of freely selected text keywords as tags, and these tags are imprecise, irrelevant, and misleading.[25] It does not follow any prescribed guidelines to assist tag formation, and tags are attached to resources based on the idea of users. Hence, a tag recommender system may provide better tag suggestions in which users select tags for a specific resource.

Features of Tag Recommending Systems
Tag recommender systems recommend relevant tags for a user provided untagged resource and are classified into personalized and collaborative tag recommendation.[72] The relevance of the tags is determined by social (collaborative) experts of the corresponding domain and by the individual users who owns the resources. Personalized recommendation approach helps individual users to comment on their content in order to govern and retrieve their own resources. Collaborative (collective) tag suggestion aims at making resources more noticeable by other users by endorsing tags that facilitate browsing. The three main features of tag recommendation system are provided in the subsequent subsections.[26]

Generality
Generality is defined as the specific characteristics of a personal or social character of the post in collaborative tagging system. The unique characteristics of a tag

are mainly used to make decision about the resources and for designing the tag recommendation system. The manual recommendation system limits the characteristics of tags according to the parameters of recommendation. Whereas the automatic recommendation system, the tags efficiently adapt to the system parameters. Therefore, optimization of the results is done by adopting efficient learning algorithm to perform automatic parameter tuning in.[81]

Adaptability

The tag recommendation process is performed dynamically depending on the user-preferred tags and the on the tags followed by other users. The continuously varying feedback from the user adds new information to the recommendation process, which can dynamically update the resources according to the content suggested by the user through post, user profiles, and associations existing between words and the preferred tags. This adaptability feature of tag recommending system facilitates online content adaptation. It improves the quality of tags extracted from a user profile and dynamically adapts it to the interest of the user.

Efficiency

Tag recommendation system is highly appreciated for its efficiency in managing a large amount of information in user-created repositories. It can produce precise results even when unlimited vocabulary of tags is available. Furthermore, tag recommendation system is extended to the collaborative tagging system and can operate even with limited resources as additional cache layer is deployed with text indexing engine.[27]

TYPES OF TAG RECOMMENDATION SYSTEMS

Tag recommendation systems are categorized based on content, collaborative, hybrid, and Flickr. Content-based techniques exclusively depend on textual metadata that is associated with the resource. Collaborative-based recommendation system mostly deals with the collaborative filtering method.[28] Hybrid systems integrate the content-based system with the collaborative-based system.[66]

Content-Based Tag Recommendation Systems

In content-based tag recommendation system, tags are extracted from the content of the blogs using artificial neural network.[29] In the Gaussian framework, the word frequencies and semantic relationships are extracted from the lexical relation of WordNet. Based on the content of the resources, the multilabel classifier extracts the title and short description of the web resources and creates tags for it. Tags derived from the various topics are combined together to perform the final recommendation task. As already available tags are reused, they become the systems main disadvantage, limiting the uniqueness of the resources and the extent of use.[30]

Collaboration-Based Tag Recommendation Systems

In collaboration-based recommendation system, most popular tags are recommended based on the social (group of users) interests, knowledge, and goals within a community.[68] This system is socially inclined as the tags are recommended on considering the opinion of the people on web resources.[31] In this approach, the user profile is constructed for each user according to their interest and the similarities between the users also computed. It deploys Singular Value Decomposition (SVD) technique to derive the relation between the user, web resources, and their corresponding tags. A probability value is assigned based to the relationship between the user and their accessing resources. As a result, the most probable tags are returned as per the user requirement. This type of recommendation systems is further categorized into model-based and memory-based approaches. In the model-based approach, the probabilistic model is constructed to predict the similarity between the users and their future rating assignments, which is built on the user browsing history. In memory-based approach, statistical techniques identify the users (neighbor) with similar behavior and recommend a list of tags from the user feedback.[32]

Hybrid Tag Recommendation Systems

The hybrid tag recommendation system produces efficient processing of the wide variety of posts by combining the strengths of the content-based and collaborative-based recommendation approaches and by overcoming their limitations.[33,34] In this approach, tags are extracted from the web resources, and there profiles are maintained by the user. Extracted tags are extended through Natural Language Processing (NLP) techniques and are combined with the tags that are extracted from a content-based approach. The similarity score, which is the measure of content of presently posted tags, is computed on the content of tags. Tag recommendation is constructed using the title of the resources, wherein the scalability of search engine and usage of source tags are important four strategies that could be used in this recommendation

process are weighted, cascade, switching, and feature combination hybrid recommender systems.

Weighted Hybrid Recommender System
A tag with a high score is recommended to the blog users, and the score for the resource is computed from the weighted sum of scores calculated by different recommender systems.

Cascade Hybrid Recommender System
This system filters a starting set of resources using a first recommender system. The ranking of these resources is then refined by a second recommender system.[83]

A Switching Hybrid Recommender System
In this system, tags are recommended based on the specific predefined criteria.

Feature Combination Hybrid Recommender System
In this system, the features of collaborative and content-based approaches are combined to perform recommendation by considering the features of collaborative as data and then context-based recommendation is performed.[32]

Flickr Tag Recommendation Systems
The Flickr tag recommendation is used to assign tags to multimedia objects.[35] Flickr is a community network that allows sharing photos with friends, family members, and so on within the network. Assigning tags to multimedia objects are difficult as no content is available to use as a resource. Therefore, the objects are tagged depending on the perception of the user with regard to the context of the object. Different users can post a large number of photos under the same category or subject and describe the content of photo using manual annotations that provide the semantic information and additional context. The fundamental aspect of this recommendation relies on the characteristics of Flickr tag. As users can describe the photo according to their own context and perspective, many users provide a different description for the same photo. This type of extended recommendation approach can provide the more enriched description about the uploaded multimedia objects. Similarly keyword queries assist the retrieval of photos.[84]

Tag Recommendation Based on Social Comment Network
Social comment network is often referred to as the characteristic representation of users based on their general interest and social association. Social comment network allows users to communicate with new participants and to maintain a relationship with anyone in the network. Group of members with similar interest unite together to form a community within which interest in specific topics are shared. Through this social network, users can transfer photos and videos with each other and receive comments from others. This system has two stages, namely developing comment network and computing author's prestige.[36]

Develop Comment Network
On social media 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.

Calculate User Prestige
The prestige of the user is computed according to its prominence on social networks. 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 by the number of links related to the corresponding resources. Thus, 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 a number of groups and are allowed to add tags to the photos.

Association between Social and Semantic Web
Tags are recommended based on the association of the semantic web on the social web. In the 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. However, Semantic web is guided under the control of World Wide Web Consortium (W3C) and extended from the existing 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.[65,86] 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

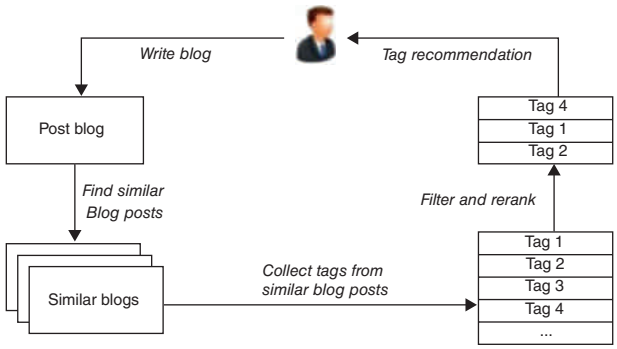

FIGURE 1 | Flow of information in AutoTag.

are recommended to the blog users. Topic ontology assists users to recommend the possible tags with a high score.

AutoTag Recommendation System

Mishne[37] presented an AutoTag is the tag recommendation methodology which uses collaborative filtering methods to suggest tags for blogs.[82,85] The valuation of AutoTag gives good accuracy on larger collection of web posts. As the final outcome, AutoTag makes the tagging process simple and helps the blogger by improving the quality of tags.

An understanding of the concepts such as "user" and "product" will give a clear idea on the process of automated tag detection. In the AutoTag mechanism, blog posts are considered as users and tags as products according to the interests of the users. Normally recommender systems has a basic concept, similar users has interests to buy similar products. AutoTag also utilizes the same concept of traditional recommender systems. AutoTag discovers high quality, useful tags for the blog post by analyzing the tags of the similar blog posts. The external knowledge of the blog posts, tags, and the preferences of the bloggers help in the further improvement of tag recommendations.

Figure 1 shows the different stages in AutoTag process. When the user creates a blog post, similar blog posts are identified. Then the tags of the similar posts are collected and the rank list of the tags is generated. AutoTag uses filtering mechanisms and reranks the tags. Finally, tags with the top-rank will be recommended to the user, so that user can select the tags for his blog post.

To determine the similarity between the blog posts, AutoTag utilizes IR methods. Huge collection posts were indexed by an IR search engine. New blog post generates query to IR search engine. Then IR engine finds the similar posts from the indexed collection of posts and highest-ranking posts are retrieved. AutoTag creates the rank list of tags from the recent and top retrieved posts using simple mechanism. Every tag generated has its own value and score in the rank list.

The prior posts and already used tags are the clear information sources available about the blogger. This information is very much useful in the filtering and re-ranking process. As an outcome, previously used tags have boosts in the rank list of the tags.

AutoTag classifiers that inherit some level of uncertainty can provide a probabilistic classification which is the limitation of the system. The predictive variance also lags in means of flexibility of making predictions to new instances.

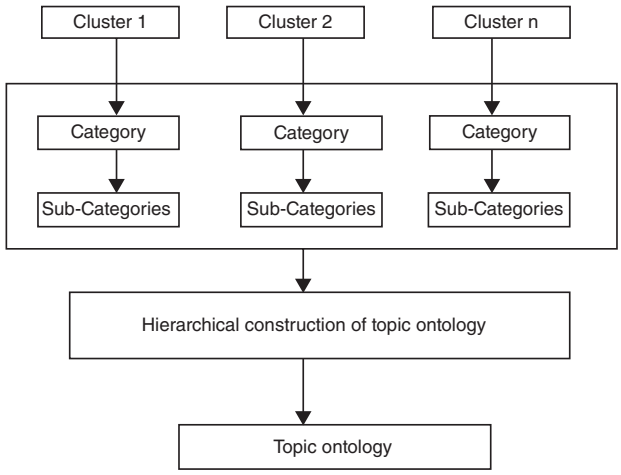

FIGURE 2 | Hierarchical construction of topic ontology.

Topic Ontology

Topic ontology is mainly used to categorize the web pages based on their content.[38] It is defined as ontology with a set of themes that are interrelated 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 resulting in the topic hierarchy. It comprises a group of relevant concepts related to the specific domain,[39] where a hierarchical semantic relationship is maintained among the concepts in the topics. Topic ontology is mainly used to categorize the content of web resources. Topic ontology is used to determine the topics for topic recommendation.[24] These topics are used as tags for the corresponding web page. Topic ontology is constructed by using the set of topics extracted from Wikipedia, and the semantic relation between the set of topics is derived from the lexical relation of WordNet.[77] Hierarchical construction of topic ontology is shown in Figure 2.

Spreading Activation Algorithm

Spreading activation, introduced by Collins and Loftus, is used for IR in computer science research and is applied in psycho linguistics and semantic priming. The spreading activation algorithm is an effective approach than most of the logic reasoning approaches as operations performed are represented in the form of pulse. Each pulse has three phases: (1) preadjustment, (2) spreading, and (3) postadjustment. Pre- and postadjustment phases manage the activation score of every node of the whole network. It monitors the activation decay level of the nodes in the network,[40] thus, preserves the node from previous pulses and avoids conflict. In the spreading phase, the activation value is spread among the nodes in the network, and the input level of the node is computed as follows:

$$I_j = \sum O_j \times W_{ij}$$

where I_j = input for the node j; O_j = output of node j which is linked to the unit level of node i; and 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 that are connected to a large number of nodes in the network. In the association IR, activation process is carried out based on the specified inference rules.[41]

Applications Overview

Based on the interest score applied on the tags, spreading activation algorithm supports spam reduction, sentiment analysis, and tag popularity by detecting spam emails, spammers, messages, web spam, and so on, consequently enhancing blog security.[79] First, spam reduction process reports irrelevant tags as spams. Second, the algorithm conducts sentiment analysis or opinion mining of the sentiments expressed by the users, which may either be positive, negative, or neutral.[80] 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. Spam and sentiment analysis is further discussed in the subsequent subsections.

PROPOSED TAG RECOMMENDATION USING SPREADING ACTIVATION ALGORITHM

In this section, we propose the use of spreading activation algorithm to study the role of constructed topic ontology for efficient tag recommendation. In the proposed approach, tags that are recommended to the user are predicted from the extracted keywords from the existing blogs and the topics in constructed topic ontology. Spreading activation algorithm determines the activation score of the tags depending on the occurrence of the terms as topics in ontology. After computing the activation score for each extracted tags from the blogs, the extracted tags that has the high activation score are recommended to the user as more relevant tags.

Spreading activation was introduced to elucidate the semantic model-based suggestions in which activation paths are stored and utilized to propose a good recommendation.[42] These paths may be exploited to produce both verbal details and act as a feedback form. Recommender systems support the user's selection process by concealing unrelated information and providing only user preferred information. Thus, recommender systems consistently provide tailored information in which user is interested. Spreading activation technique identifies the topics of interest to the users, runs and spreads activation power to all interrelated concepts and items and assigns weights on iteration. This approach is described by the branch and bound approach. Furthermore, link weight prior to the execution of spreading activation execution, network should start prior to the commencement of spreading activation execution. In addition, link weight that is set for each node in a network is assigned based on user background and activation values; however, it should be noted that network starts prior to the commencement of spreading activation execution. Activation begins with certain value that is received by initial nodes and the level of activation is computed by assigning activation. Primary nodes are added in a priority queue arranged with downward activation.[43]

On performing the initialization, the highest weighted node is removed from the priority queue and the node activation spreads to all the nearest nodes. These nearest nodes are then added into the queue if they are not labeled as processed. The node that spreads its activation to the nearest nodes is labeled as processed. After receiving suggestions, the user avails more choices to use tags and aid its easy discoverability. Spreading thus is triggered to all items and displayed in a final suggestion list. On examination of the activation paths, the accuracy of the initial node paths

could be demonstrated. Furthermore, spreading activation streams from items are avoided by setting up of the items and their associations on the stop list.[42] Table 1 shows some examples of the most popular topics with frequent tags.

Spreading activation algorithm provides the top tags (top-p) with parameters for users based on tag scores. Thus, spreading activation algorithm maximizes the tag popularity (giving a value of 1 for this parameter). Tag popularity can also be the called weight of the tags which is computed by applying interest scores on the tag occurrences. The most popular tags are secured to most resources. Based on the occurrence of tags in a blog post, tags are used as recommendations or suggestions. Interest scores (tag weights) are automatically modified based on the occurrences of tags and activation levels in topic ontology.[25,44,45]

TABLE 1 | Most Popular Topic with Frequent Tags

Topic	Frequent Tags
Download	bittorrent, p2p, torrent, torrents, latex, anime, tex
Programming	Microsoft, C++, C, dotnet, c#, JAVA
Social Network	social, community, collaboration, networks, network, research, web, social software
Security	hacking, hack, wifi, hacks, firewall, wireless
Entertainment	Shopping, Cinema, Playing, fashion
Awards	Nobel, Oscar, Miss world, Academy
Sports	Cricket, Football, Tennis, Golf, Olympics
Design	web design, portfolio, graphic, portal, graphics, illustration, graphic design
Scripting	html, web, standards, xhtml, accessibility, w3c, web design,
News	daily, magazine, imported, magazines, media, newspaper, newspapers, TV

An algorithm of Spreading Activation

```
Input: Interest score and tags of existing blogs
Output: Recommendations with updated tags
Tags = {T1, ...,Tn}, tags with interest scores
Interest score (Ti),
Interest score (Ti) =1, no interested tags
I = {b1, ..., bn}, blogs
for each bi Є I do
 Initialize queue;
 for each Ti Є Tags do
        Ti. Activation =0;
 end
 for each bi Є I do
        Compute sim (bi, Ti);
          If sim (bi, Ti) > 0 then
                Ti. Activation = Interest score (Ti) * sim (bi, Ti);
                Queue. Add (Ti);
          else
                Ti. Activation =0;
          end
 end
 While Queue. Score > 0 do
        Order queue // Activation values (Ascending)
        Ts= Queue [0]
        If pass limitations (Ts) then
                Relevant Tags= Get Relevant Tags (Ts);
        for each TR in Related Tags do
                TR. Activation += TS. Activation * TR. Weight;
        Queue. Add (TR);
        end
        end
 end
end
```

Spreading activation is a cognition model that describes the psychological result of perception and logic reasoning. It is more useful in semantic networks related computational research area in the information retrieval. In spreading activation, all related information is considered as node and is plotted on a graph with activation level. Relations among the concepts in spreading activation are denoted using link between nodes. Nodes in a graph begin their activation level and spread it to the nearest and relevant nodes. In this approach, most relevant tags are captured from the constructed topic ontology using the spreading activation algorithm. It is mainly used to identify the similarity between the keywords in a blog and concepts in constructed topic ontology. Extracted keywords can be recommended to the user as tags to the blog user in a particular context. Similarity is computed based on the activation flow of originally activated nodes. Many nodes are triggered to a particular degree, i.e., associated to initially chosen concepts. It is possible to make use of such spreading activation to recommend tags by activating interest scores to the nodes. Initially, activation value is set to zero (0). If the user clicks on a specific, it is understood that beginning node is activated to the stream of activation. As a consequence, topics and tags are relevant to the recent one that contains highest activation value. If user is interested in a particular tag and clicks on the related link, system introduces primary activation into the nodes in a network representing initial tag name. Thus, activation is spread to semantically correlated nodes as well as spread the sum of received activation to other nodes. Finally, all nodes may contain certain activation value that denotes their relevance degree.

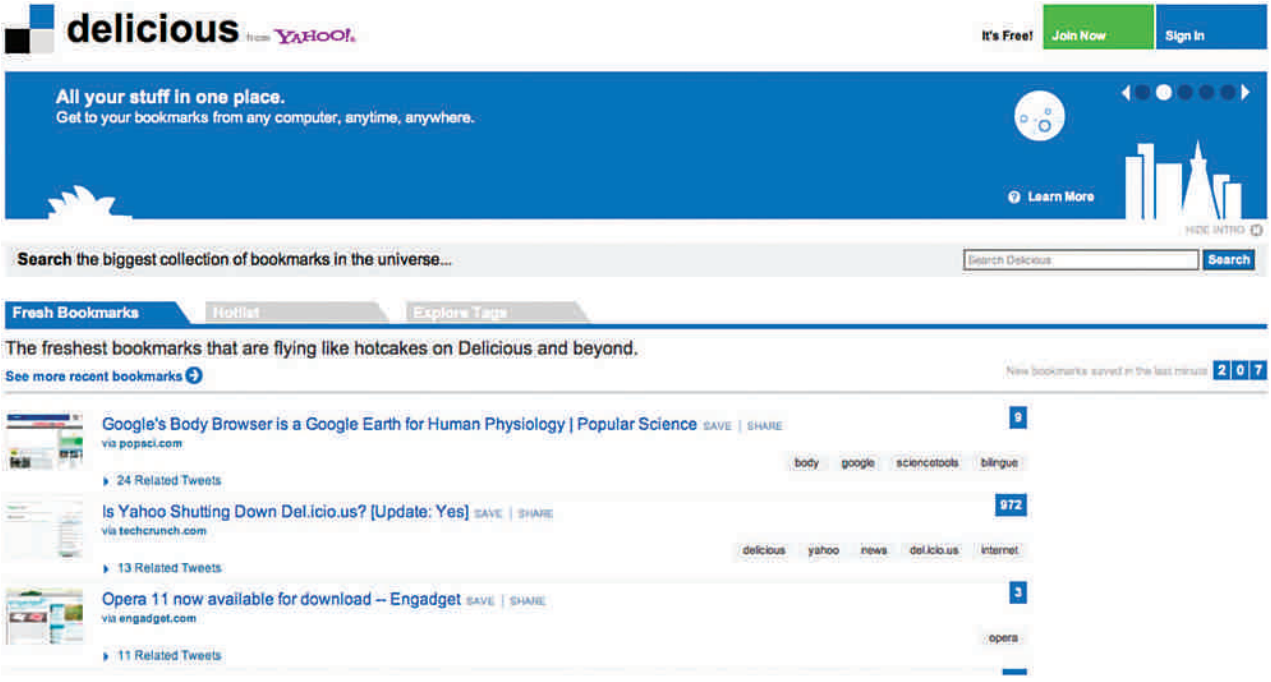

FIGURE 3 | Snapshot of Delicious.

The method of applying topic ontology along with spreading activation algorithm will produce an efficient recommendation such as popular tag recommendation for a particular resource. Interest scores were calculated and assigned to the blog content and tags which is already extracted. Every extracted tag was assigned with weights. These tag weights are calculated on the basis of activation score. This activation score is assigned by the similarity index between the content of blogs and topics of the constructed ontology.[46] The quality of the tags was determined by highest activation score. The tags with high activation score is recommended to the users. The tag recommendation method is evaluated experimentally using real-world large datasets. The capabilities of topic ontology with spreading activation algorithm for the efficient tag recommendation are compared with the AutoTag method through the experimental results.

Experimental Setup

In this research, the dataset used for experimental setup was collected from Folksonomy-oriented bookmarking sites.[76] The main objective of this research was to provide an effective tag recommendation for the resource. Based on the blog contents, Folksonomy dataset collection was used to find the relevant tags in the present experiments. Common metrics, such as precision, recall, and *F*-measure, were selected as benchmark of the obtained results. For each of the data to be tested, the tags were extracted from blogs using keyword extraction method and interest scores for keywords were computed to set up the input for recommendation. The datasets, such as BibSonomy and Delicious, are examples of extracted tags from the blogs.

Datasets

The selected dataset consisted of tags, titles, category from Wikipedia, and semantic relationship from WordNet. The main objective of the selection phase was to construct topic ontology related to tags. Existing blog tags were used as test datasets. Furthermore, for the purpose of evaluating the proposed recommendation approach, datasets were chosen from two different Folksonomy systems, namely Delicious and BibSonomy. These are popular social networking systems that provide interesting recommendations to the community, especially the research fraternity, as they permit users to convey their thoughts on resources with their own words.

Delicious

Delicious datasets can be used for a limited period of time in which user can build bookmark to the URL and share with others. Delicious is one of the popular collaborative tagging sites for bookmarking that permit users to tag blog and web pages on the web. Figure 3 illustrates a snapshot of Delicious.

TABLE 2 | Most-Frequented Domains in the Delicious Corpus

Domain	Bookmarks	Users
en.wikipedia.org	937,785	305,739
www.flickr.com	892,157	262,963
www.youtube.com	890,769	256,126
www.google.com	772,460	176,890
www.nytimes.com	613, 676	121, 575
www.amazon.com	541, 314	94,093
news.bbc.co.uk	416,878	85,910
lifehacker.com	369,078	80,728
community.livejournal.com	320, 021	39,755
www.microsoft.com	310, 701	131,847

TABLE 3 | Top 10 Popular URLs in the Delicious Corpus

URL	Bookmarks
www.flickr.com	35,732
www.pandora.com	35,531
script.aculo.us	31,643
www.netvibes.com	30,782
en.wikipedia.org	27,672
www.youtube.com	26,183
slashdot.org	25,630
www.last.fm	23,957
oswd.org	21,530
www.alvit.de	21,130

Table 2 shows the most frequented domains in the Delicious corpus. Tags can be added to the users' bookmark to explain, search, share, and classify the bookmarks. Statistics on the most recent bookmarks and its corresponding tags are shown in Delicious' front page. Delicious also has a popular page to view the same information for most popular URLs. A set of 10 tags were considered for this research work. The frequently occurring tags and the most popular tags were assessed from 23,701 URLs. Of the 2,01,711 tags that were retrieved, we found that 89% of tags were obtained for each URL. Thus, it can be understood that finding relevant topics is not difficult as most of the users tag the content. Table 3 shows the most frequently visited popular URLs in the Delicious corpus. Figure 4 shows the popular URLs in Delicious corpus.

BibSonomy

BibSonomy is possibly the best investigated Folksonomy to date in which user can accumulate and

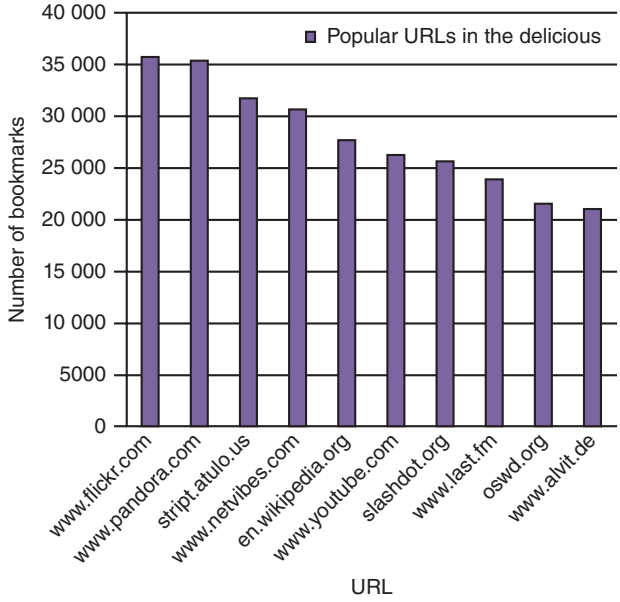

FIGURE 4 | Popular URLs in the Delicious corpus.

interpret URLs and publications as well. Bibsonomy dataset is employed for tag recommendation challenge. In this research, users, resources, tags, or keywords are considered as datasets while other additional data have been disregarded or ignored for all practical purposes. A set of 10 tags were chosen randomly from the tag list, and the bookmark content were retrieved for each tag with respect to relevant tags. Figure 5 shows the snapshot of BibSonomy.

Results and Discussion with Bibsonomy and Delicious

Popular tags provide an accurate and valuable retrieval results. The tags of BibSonomy and Delicious provide most popular tags with high interest scores. Tags are important in ranking multiple keywords and relating the blog post to the appropriate tags. From users' perspective, use of most popular tags is fine grained for recommendation process. Delicious page provides bookmarks in reverse sequential order and tags that user assigned to the bookmarks as any one can filter the list of bookmark of anyone's.

Delicious allows users to tag the resources uploaded by others. The primary function of Delicious is to suggest interested websites and people interested in the resources to the user. Table 4 represents top 15 most popular tags in BibSonomy and Delicious with their interest scores. Figures 6 and 7 show the interest scores for BibSonomy and Delicious.

Table 5 shows top 15 frequent tags with their interest score in the corpus and Figure 8 shows the frequent tags with their interest score.

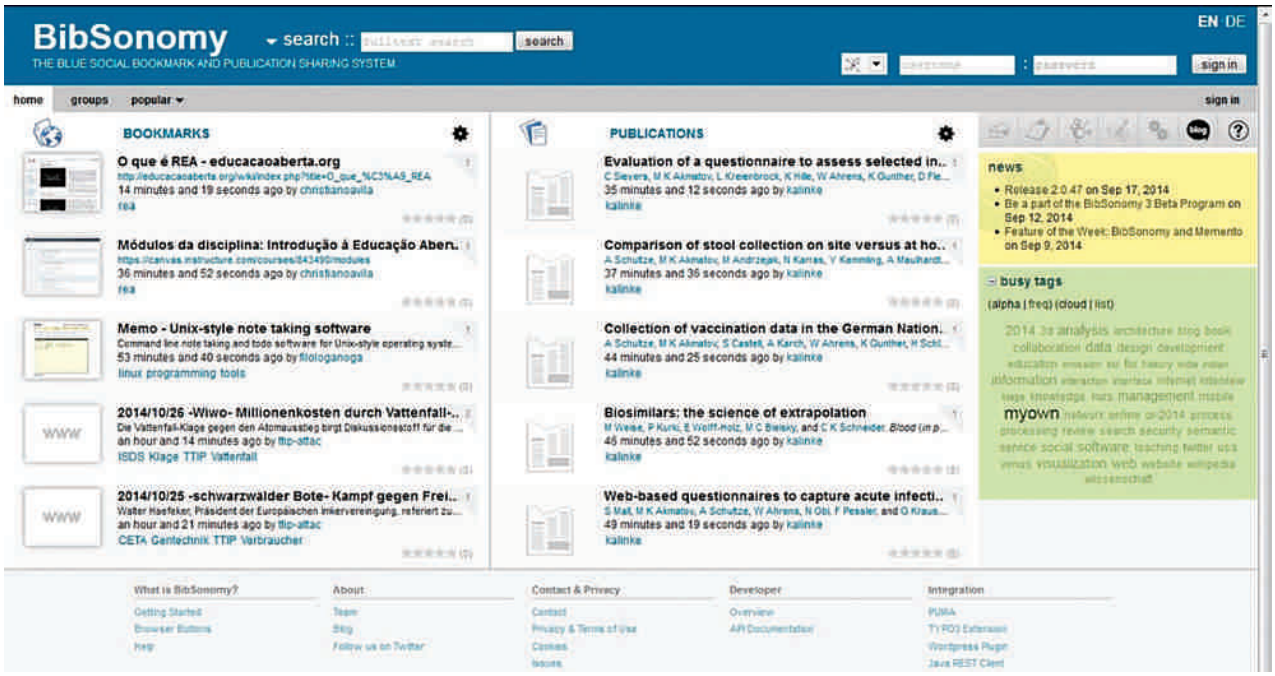

FIGURE 5 | Snapshot of BibSonomy.

TABLE 4 | Top 15 Most Popular Tags in BibSonomy and Delicious with their Interest Scores

BibSonomy		Delicious	
Tags	Interest Scores	Tags	Interest Scores
Art	1302	Computing	2640
Sports	2281	Internet	1484
About Me	1500	Schools	1098
Awards	1248	Software	2495
Academic	2253	Academic	1974
Tips	2326	Books	1717
Books	2456	Networking	118
Software	1974	Web	178
Sceneries	1328	Technologies	1258
Web	297	Music	1196
Schools	1233	News	1015
Music	2645	Awards	963
Photos	1429	Travels	1375
News	861	Interview tips	1212
Cosmetics	1258	Cartoons	1634

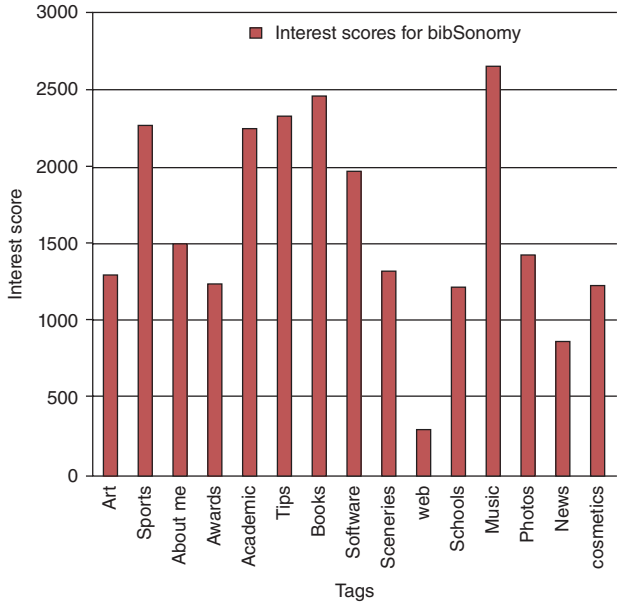

FIGURE 6 | Interest scores for BibSonomy.

Performance Evaluation Metrics

This section presents the three most widely used metrics for evaluating the performance of recommendations produced by a recommender system. These standard information retrieval metrics include Precision, Recall, and *F*-measure.[47] A user expects an ordered list of recommendation, perhaps from best to worst, as the end product from a recommender system. In many cases, the users may not be particular about the order but would be able to appreciate few good recommendations in any order. Therefore, the evaluation of the recommendation system should be made on how information is retrieved on these systems using

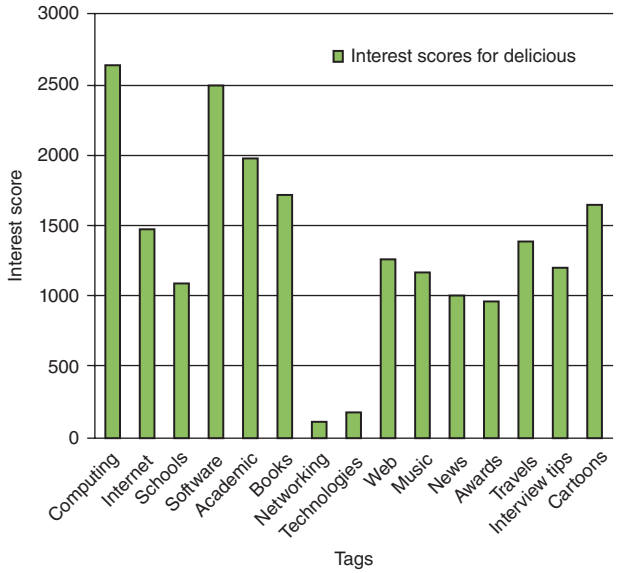

FIGURE 7 | Interest scores for Delicious.

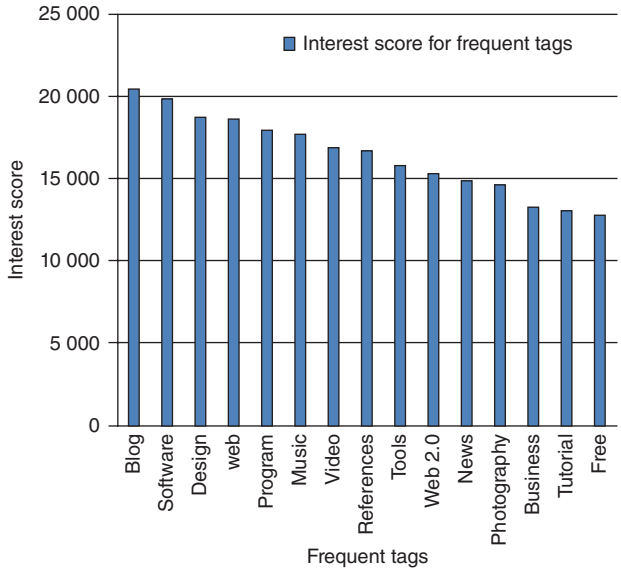

FIGURE 8 | Interest score for frequent tags in corpus.

TABLE 5 | Top 15 Frequent Tags with their Interest Score in Corpus

S. No.	Tags	Interest Score
1	Blog	20,543
2	Software	19,847
3	Design	18,745
4	Web	18,621
5	Program	17,923
6	Music	17,612
7	Video	16,821
8	References	16,567
9	Tools	15,734
10	Web 2.0	15,142
11	News	14,873
12	Photography	14,564
13	Business	13,191
14	Tutorial	13,012
15	Free	12,745

precision and recall. These metrics are often applied to search engines which produce some of the best outcomes on the query raised. Furthermore, these metrics when used in combination proves to be an effective measure to evaluate the recommendation system.[48–50]

Precision

In IR, precision is the portion of retrieved instances that are relevant and measures the quality of the recommended tags. It is the number of relevant tags retrieved versus the total number of retrieved tags. Thus, precision can be referred to as the proportion of the correctly retrieved tags from all the retrieved tags by the recommendation system can be measured.[51]

$$\text{Precision} = \frac{|\text{relevant tags}| \cap |\text{retrieved tags}|}{|\text{retrieved tags}|}$$

Recall

In IR, recall is the portion of relevant instances that are retrieved and measures the completeness of the recommended tags. Thus, recall is defined as the number of relevant tags (e.g., instances belonging to a particular relevant category) retrieved divided by the total number of tags available in all the relevant documents.[51,52]

$$\text{Recall} = \frac{|\text{relevant tags}| \cap |\text{retrieved tags}|}{|\text{relevant tags}|}$$

F-Measure

F-Measure combines recall and precision into one measure for comparison, and it inclines toward the smaller of the two values. Thus,

$$\textit{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

It is often referred to as *F*1 measure, as precision and recall are weighted equally.

Comparison and Performance Evaluation

In this research, we have made an attempt to compare the tags generated by the spreading activation system with the existing AutoTag mechanism. The performance of the proposed approach, i.e., spreading

TABLE 6 | *F*-Measure of Both BibSonomy and Delicious Datasets

	BibSonomy		Delicious	
No. Tags	AutoTag	Spreading Activation	AutoTag	Spreading Activation
1	0.15	0.15	0.13	0.16
2	0.17	0.20	0.16	0.18
3	0.19	0.26	0.20	0.26
4	0.23	0.30	0.23	0.30
5	0.26	0.33	0.26	0.33
6	0.29	0.36	0.28	0.35
7	0.31	0.40	0.30	0.37
8	0.34	0.41	0.31	0.39
9	0.37	0.44	0.32	0.40
10	0.37	0.46	0.34	0.43

activation system, was measured based on the collected data from two different Folksonomy systems. Folksonomy is a social and decentralized approach that is formed by individuals or groups. Existing AutoTag mechanism does not recommend newly added tags when it is used already in a blog (Table 6). Figures 9 and 10 illustrate the *F*-measure for the BibSonomy and Delicious datasets, respectively.

The obtained results demonstrate the superiority of the proposed approach with respect to the performance. In comparison with AutoTag mechanism, it can be understood that the datasets generated by spreading activation system gradually increases when more tags of the recommendation is used in Delicious datasets. Thus, the proposed algorithm achieves higher performance than the existing AutoTag mechanism on the tags; however, it is much difficult to identify the resource specific to the most popular tags. Although, the proposed approach identifies the semantics of tags and resources, the approach of discovering semantics varies from the AutoTag mechanism. The detailed dataset holds essential metrics and plots and so it provides better results.

The results of the experiment clearly demonstrate that the proposed method was significantly effective for the most frequent tag search, as against the result obtained for all tags. Furthermore, the proposed method showed largest improvement in the top few tags. However, the accuracy of the recommendation decreases with the increasing size of the most frequent tags set, which is an expected behavior, given that less frequent tags would become harder to recommend. The same pattern can be observed for user relevant tags, which shows that the spreading activation is not impairing the quality of recommendation for high-frequency tags.[53]

The proposed tag recommendation system is compared with the existing approaches such as Autotag and spread activation by means of comparing the performances. Existing mechanisms do not provide any new tags. They concentrate on using already used tags of a blogs. The proposed model has improved a lot by providing new tags as suggestion and it is

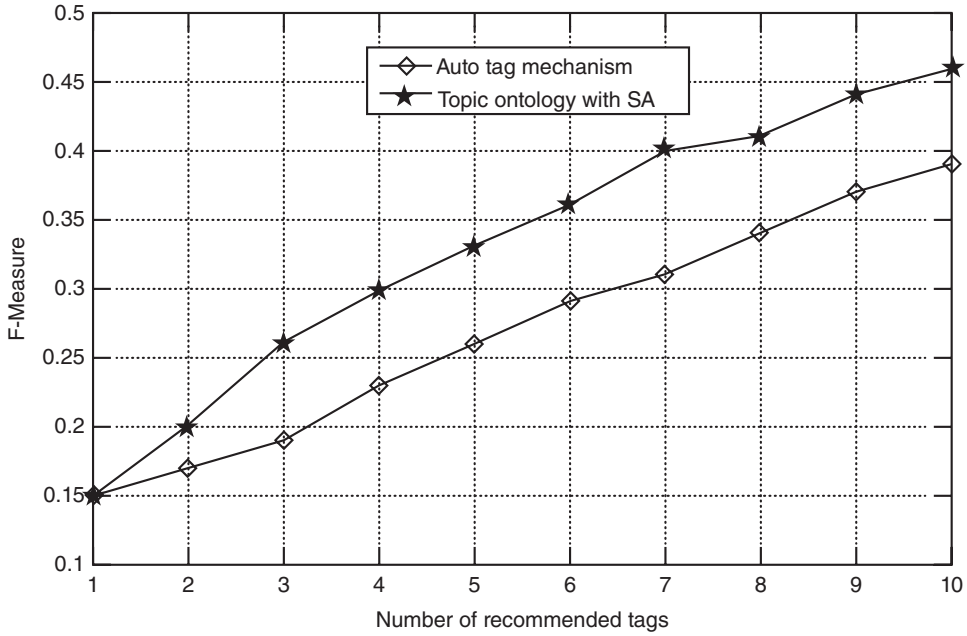

FIGURE 9 | *F*-Measure for BibSonomy datasets.

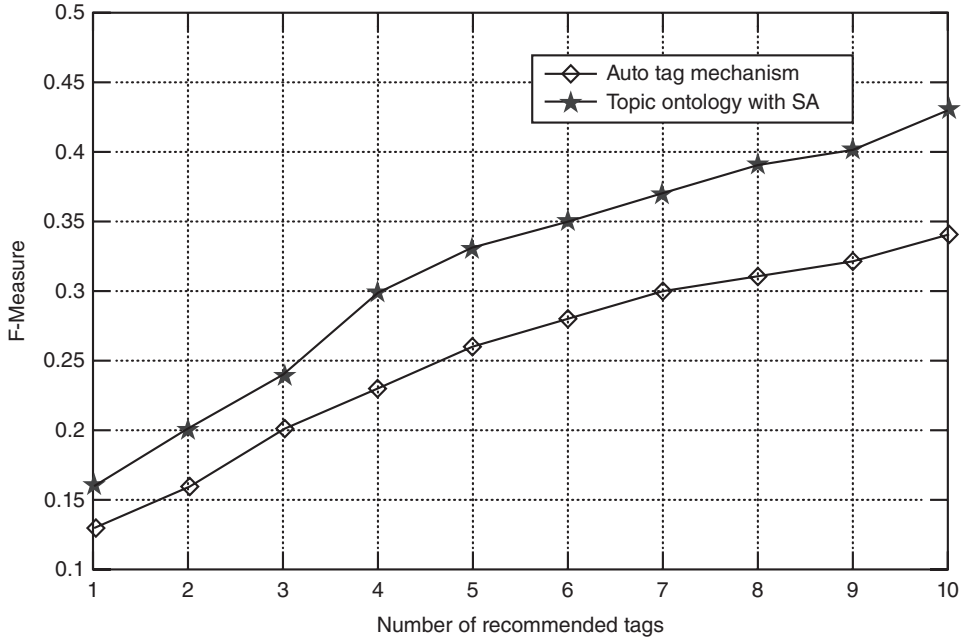

FIGURE 10 | *F*-Measure for Delicious datasets.

proved by its precision percentage too. The proposed tag generation system out performs by proving itself through providing high accuracy. In the generated list of tags, top tags are more related than the tags which are in the bottom of the list. For the most popular tags, it is tedious process to rank through using resources of the blog information. Here the proposed method performs better than the existing methodologies. The approach of discovering semantics varies from the existing mechanisms and provides better results.

PROPOSED TAG CLASSIFICATION SYSTEM

In the proposed tag classification system, the features are extracted from CiteULike dataset. A feature expansion using hypernym and hyponym and Correlation-Based Feature Selection–Hybrid Genetic Algorithm (CFS-HGA) is proposed and compared with other existing feature selection methods. Also, a classifier HGA-SVM (support vector machine) is proposed.

CiteULike Dataset

CiteULike dataset is a free online social bookmarking service that allows researchers to share, store, and organize information about the scholarly papers. Links are added by users and references are imported from other scholarly digital libraries available on CiteULike. For instance, users can link to a paper in CiteSeer or ACM through their personal collection on CiteULike. This dataset can also provide additional information, such as the BibTeX entry and all the tags included for that paper and Ref. 54.

In the CiteULike dataset, of the 32,242 tag applications available, there are 2011 distinct users, 9623 distinct papers, and 6527 distinct tags. The two most active users alone have 3883 and 634 tag applications, while 42 users have 100 or more tag applications. The two most tagged papers have been coauthored by Larry Page and each being tagged 135 and 94 times, respectively. The five most popular tags are clustering, p2p, logic, learning, and network. The average number of tag applications per paper is found to be 3.35 which is calculated as the total number of tag applications divided by total number of papers. As the distribution of tag applications per paper was skewed, the median and modal numbers of tag applications are considered and were found to represent a more realistic picture of tagging behavior where the median and modal numbers of tag applications per paper are 2 and 1, respectively.

Feature Selection

Term Frequency–Inverse Document Frequency

Term Frequency–Inverse Document Frequency (TF-IDF) is obtained from TF × IDF. The TF is called the word frequency that is it refers to the frequency of terms if the exclusion of forbidden and individual high-frequency words, the more frequency of terms

in the given document, the stronger ability of terms for the document characterization.[55] TF-IDF is also adjusted for number of records (or documents) having that word. The TF-IDF formula can be represented as:

$$\text{TF-IDF} = \frac{\text{Frequency}\,(i) * N}{\text{d}f\,(i)}$$

where df is the frequency of word (i) in all documents, N, the number of words in the record/document, and i, the word list in record/document.

In the case of TF-IDF, term frequency for every word can be normalized by IDF which reduces weight of terms occurring frequently in a collection.[56,70] This also reduces the importance of common terms in a collection, ensuring that document matching is influenced by more discriminative words with relatively low frequencies in a collection.

Feature Expansion Using Hypernym and Hyponym

A term/phrase can have many meanings, while a domain-specific concept is unambiguous. Ontology learning implies extraction of conceptual knowledge from many sources to build ontology from scratch. Enriching/adapting existing ontology is an attempt at knowledge acquisition.

Hyponym are specific instantiations of a more general concept, e.g., red and color. Thus, the hyponym word provides a more specific type of concept than when it is displayed by the other. Proper nouns are examples of hyponyms. Hypernyms refer to broad categories/general concepts. Color or flower is the example of hypernyms for precise terms such as blue or rose.[57]

In the proposed method, the search can be broadened by the use of hypernyms. It uses the domain-specific concepts in documents than terms to retrieve documents from a specific domain. Thus, the list of hyponym present in documents is extracted. More than one term may refer to the same concept in some cases. The concept frequency includes frequencies of the synonymous terms of the concept in the document.[58]

A list of terms and frequencies exists for each document where an associated set of hyponym for each term is obtained. A term maps to one or more hypernym. For example, the term 'structure' can map to buildings, physics or anatomy. Of the mapped hyponym, the most appropriate domain is located. A hyponym becomes significant when the document consists of many related concepts of that particular term. The proposed algorithm uses document terms with their frequency as input, returning a concept list along with the significance of the document.

In the proposed feature expansion using hyponyms, for each term t_i in the term list of a document D, the hyponyms h_{ij} are obtained. Let the impact of each associated concept h_{ij} be hi_{ij}. The impact hi_{ij} is assumed to be the normalized frequency of the term t_i, i.e., t_i *frequency*. For each associated hyponym h_{ij}, the presence of the related hypernym h_r in the document is considered. The impact of the associated hyponym h_{ij} is then incremented by α* normalized term frequency for the occurrences of the terms t_p corresponding to the hypernym h_r.

$$hi_{ij} = t_i \text{ frequency } + \alpha * t_p \text{ frequency}$$

where α is the weight and in this study, the value given is $\alpha = 1/2$

Support Vector Machines

The major strength of SVM is a rather easy training. SVMs scale up relatively well- to high-dimensional data and the tradeoff between classifier complexity and error can be controlled explicitly.

Radial Basis Function

The classification problem cast into a higher dimensional space becomes more likely separable than in a lower dimensional space. Radial basis function (RBF) can be used find a set weights for such problem. The weights are located in the higher dimensional space than the original data.

Finding a surface in the higher dimensional space is realized through learning. The best fit to the training data is provided by this learning. RBFs were made available by the hidden layers, which will represent an arbitrary basis for the input patterns during the expansion of hidden space.

It is a real-valued function whose value depends on the distance from the origin and any function ϕ that satisfies the property $\phi(\text{r}) = \phi\,(||\text{r}||)$ is radial function. RBFs are also used as a kernel in support vector classification. The summing up of the RBFs results in approximating the given functions. This approximation can be taken as a plain kind of neural network.

There are three basic classes of RBFs:

1. Multiquadrics:

$$\Phi\,(r) = \left(r^2 + c^2\right)^{1/2} \text{ for some } c > 0 \text{ and } r \in R$$

2. Inverse multiquadrics:

$$\Phi\,(r) = \frac{1}{\left(r^2 + c^2\right)^{1/2}} \text{for some } c > 0 \text{ and } r \in R$$

3. Gaussian functions:

$$\Phi(r) = \exp\left(\frac{r^2}{2\sigma^2}\right) \text{ for some } \sigma > 0 \text{ and } r \in R$$

Gaussian functions are probably the most used. In general, the selection depends on the application.

SVM PolyKernal

In machine learning, the polynomial kernel is a kernel function which is commonly used with SVMs and other kernelized models that represent the similarity of training samples in a feature space over polynomials of the original variables, thus allowing us to learn nonlinear models.

Naturally, the polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations some of the features. In terms with regression analysis such combinations are known as interaction features. The hidden feature space of a polynomial kernel is equivalent to that of polynomial regression.

SVM is used in pattern recognition task. Input data are transformed from input space with lower dimension m_0 to a feature space with higher dimension m_1. The separating hyperplane, i.e., weights w_j are estimated in the feature space. The optimal hyperplane is constructed in the feature space using the innerproduct kernel without considering the explicit form of feature space. Inner product kernel loosely speaking: It is a function replacing activation function of a single neuron network.

1. SVM classification:

$$\min_{f,\xi_i} \|f\|_K^2 + C\sum_{i=1}^{l} \xi_i \; y_i f(x_i) \geq 1 - \xi_i, \quad \text{for all } I;\ \xi_i \geq 0$$

2. SVM classification, Dual formulation:

$$\min_{\alpha_i} \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$0 \leq \alpha_i \leq \mathrm{C}, \text{for all } i; \sum_{i=1}^{l} \alpha_i y_i = 0$$

Variables ξ_i, called slack variables, express a measure the error made at point $(\mathbf{x}_i, y_i)$. Training SVM becomes quite challenging when the number of training points is large.

Training Set

Given a training set of N data points $\{y_k, x_k\}_{k=1}^{N}$, where $x_k \in R^n$ is the kth input pattern and $y_k \in R$ is the kth output pattern, the classifier can be constructed using the support vector method in the form

$$y(x) = \text{sign}\left[\sum_{k=1}^{N} \alpha_k y_k K(x, x_k) + b\right]$$

where α_k are called support values and b is a constant. The $K(\cdot,\cdot)$ is the kernel, which can be either $K(x, x_k) = x_k^T x$ (linear SVM); $K(x, x_k) = \left(x_k^T x + 1\right)^d$ (polynomial SVM of degree d); $K(x, x_k) = \tan h\left[\kappa x_k^T x + \theta\right]$ (multilayer perceptron SVM), or $K(x, x_k) = \exp\left\{-\|x - x_k\|_2^2 / \sigma^2\right\}$ (RBF SVM), where κ, θ, and σ are constants.

Correlation-Based Feature Selection–Hybrid Genetic Algorithm

CFS evaluates the subset of attributes by weighing the individual predictive ability of feature together with the degree of redundancy between them.[59] The correlation between attribute subset and class, and the intercorrelations between the features is estimated by correlation coefficients. The weight of a attribute subset increases with the correlation between features and classes, and decreases with growing inter-correlation. Thus, CFS is ideal for determining the best feature subset and is frequently combined with search strategies, such as forward selection, backward elimination, bidirectional search, best-first search and genetic search. The equation for CFS is given as follows:

$$r_{zc} = \frac{k\overline{r}_{zi}}{\sqrt{k + k(k-1)\overline{r}_{ii}}}$$

where r_{zc} is the correlation between the feature subsets and the class, k is the number of subset features, r_{zi} is the average of the correlations between the subset features, and the class r_{ii} is the average intercorrelation between subset features.[60]

Genetic Algorithm Feature Selection

Genetic Algorithm (GA) is used as a global search method capable of exploring large search spaces effectively. A GA is composed of three operators: reproduction, crossover, and mutation. Reproduction is used to select good string, crossover is used to combine good strings for generating a better offspring, and mutation alters a string locally to create an enhanced string. For each generation, the fitness of the population is evaluated and checked for termination criteria. The population is further operated upon and re-evaluated if the termination criterion is not met.[61]

Correlation-Based Feature Selection–Hybrid Genetic Algorithm

The HGA-CFS process is based on the principle of the fittest member in a population, which retains its genetic information by passing it on from one generation to the other. The process of HGA-CFS can be described as follows:

- *Initialization*: Generate a random initial population of n chromosomes.
- *Fitness Evaluation*: Evaluate the fitness function $f(x)$ for each chromosome x based on the CFS.
- *Selection*: Two parent chromosomes are selected based on their fitness for reproduction.
- *Crossover*: To form a new offspring (children) from the parents using a single-point crossover probability.
- *Mutation*: Mutate the new offspring at each position using a uniform mutation probability measure.[62]

Classification with HGA-SVM

SVM technique is a machine-learning method that was introduced by Vapnik and coworkers. This method is modified for different applications such as classification, clustering, data reduction, feature extraction, and regression.[63,67] Generally SVM is a two-class classifier.[75] The training samples are mapped into a high-dimensional space and a separating hyperplane is found that maximizes the margin between two classes.[64]

The most commonly used kernels for SVM are polynomial function and RBF. RBF is found to be effective as vectors can be nonlinearly mapped to a very high-dimensional feature space. The efficiency of the RBF kernel depends upon the constants γ and C, where γ is the width of the kernel function and C is the error/tradeoff parameter. In the proposed HGA-SVM, the optima value of γ and C are obtained using GA. Each chromosome represents C and γ. The fitness function for each chromosome based on the classification accuracy is evaluated.

Results and Discussion on CFS-HGA

The proposed methods are evaluated using CiteULike dataset. An attempt has been made to evaluate the various feature selection method and classifiers for classifying the tags in this research. The features were extracted using TF-IDF, proposed feature expansion using hyponym and hypernym, CFS feature selection, CFS-GA feature selection method and CFS-HGA feature selection method. The features selected were classified using SVM classifier. The performance of the classifier with polynomial kernel, RBF kernel and the proposed HGA-SVM was evaluated. Table 7 tabulates the precision, recall, and *F*-measure achieved for the various techniques.

From Figure 11 and Table 7, it can be observed that the precision of proposed methods CFS-HGA feature selection and HGA-SVM achieves the best result of 0.7942. For TF-IDF, the proposed HGA-SVM increases precision by 2.24% than SVM-RBF kernel and by 3.93% than SVM polynomial kernel. For feature expansion, the proposed HGA-SVM increases precision by 4.94% than SVM-RBF kernel and by 7.42% than SVM polynomial kernel. For CFS-HGA feature selection, the proposed HGA-SVM improves precision by 2.44% than SVM-RBF kernel and increases by 13.98% than SVM polynomial kernel.

From Figure 12 and Table 7, it can be observed that the recall of 0.6844 is achieved for the proposed methods. For TF-IDF, the proposed HGA-SVM increases recall by 2.86% than SVM-RBF kernel and by 10.2% than SVM polynomial kernel. For feature expansion, the proposed HGA-SVM increases recall by 2.35% than SVM-RBF kernel and by 4.38% than SVM polynomial kernel. For CFS-HGA feature selection, the proposed HGA-SVM increases recall by 4.31% than SVM-RBF kernel and increases by 28.45% than SVM polynomial kernel.

From Figure 13 and Table 7, it is observed that the *F*-measure is calculated for different classifiers and proposed method achieves the best result of 0.735. For TF-IDF, the proposed HGA-SVM increases by 2.54% than SVM-RBF kernel and increases by 7.08% than SVM polynomial kernel. For feature expansion, the proposed HGA-SVM increases by 3.7% than SVM-RBF kernel and by 5.96% than SVM polynomial kernel. For CFS-HGA feature selection, the proposed HGA-SVM increases by 3.38% than SVM-RBF kernel and by 21.69% than SVM polynomial kernel.

In the proposed tag classification system, feature expansion is done using hypernym & hyponym through CFS-HGA. To classify tags, HGA-SVM is introduced as a classifier. TF-IDF is used to exclude the forbidden and individual high-frequency words terms in the given document. Table 7 shows that the proposed tag classification system (CFS-HGA) using classifier HGA-SVM outperforms the other systems in almost all aspects, especially with respect to extracting features and classification. As overall, the system can extract, link, classify, and tag with high-performance ratio. For tag classification, the data are collected from citeUlike. Tag classification schemes using HGA-SVM as classifier are used to generate graphical output for performance evaluation. The generated precision, recall, and *F*-measure show the performance

TABLE 7 | Different Classifiers for Precision, Recall, and *F*-Measures

	TF-IDF	Feature Expansion	CFS Feature Selection	GA Feature Selection	CFS-HGA Feature Selection
Precision					
SVM Polynomial kernel	0.4736	0.5146	0.4942	0.5838	0.6968
SVM-RBF kernel	0.4814	0.5268	0.5655	0.5924	0.7753
HGA-SVM classifier	0.4922	0.5528	0.5744	0.6959	0.7942
Recall					
SVM Polynomial kernel	0.4314	0.4926	0.5012	0.5492	0.5328
SVM-RBF kernel	0.4622	0.5024	0.5224	0.5519	0.6561
HGA-SVM classifier	0.4754	0.5142	0.5362	0.5968	0.6844
F-Measure					
SVM Polynomial kernel	0.452	0.503	0.498	0.566	0.604
SVM-RBF kernel	0.472	0.514	0.543	0.571	0.711
HGA-SVM classifier	0.484	0.533	0.555	0.643	0.735

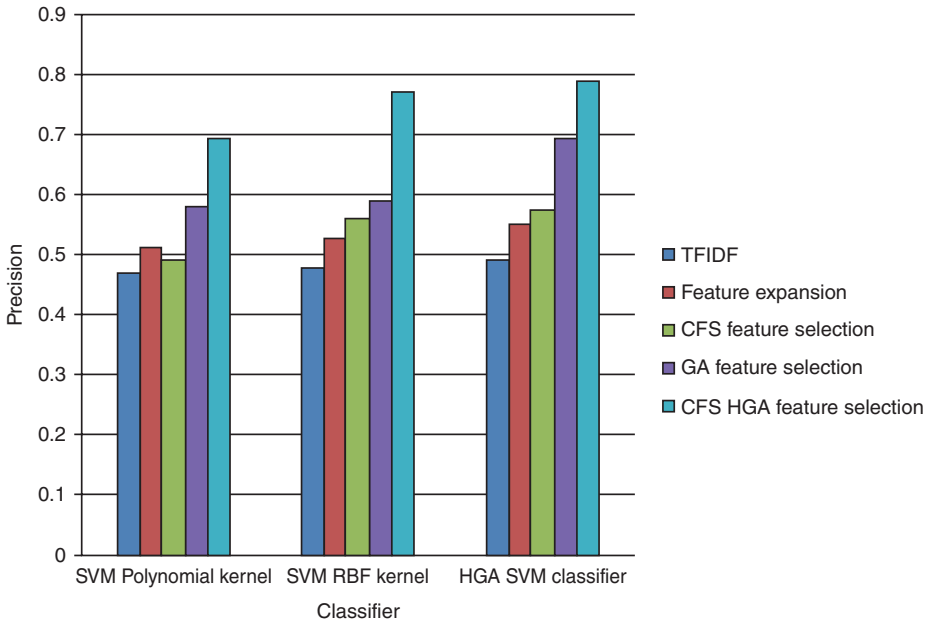

FIGURE 11 | SVM PolyKernal, RBF Kernal, and HGA-SVM versus precision comparison.

of the classifier and classification system.[71] Using weight scheme TF-IDF for tag classification improves accuracy.

CONCLUSION

With widespread research in a large number of areas, enough resources are available on the web. Overwhelming information availability blurs the visibility of specific information sought for. This article has presented an effective approach to sieve the appropriate resource from an abundant available source. To achieve this goal, an efficient tag recommendation system that recommends the most appropriate tags to the user using semantic ontologies constructed from the content of existing blogs is necessary. The issues presented in this article will further advance the discussion on producing next generation technologies with improved tag recommendation on the social tagging systems. The use of data mining algorithms has changed the types of recommendations as applications progress from recommending what to ingest to recommending when to ingest. While recommender systems have started as a transient innovation, they

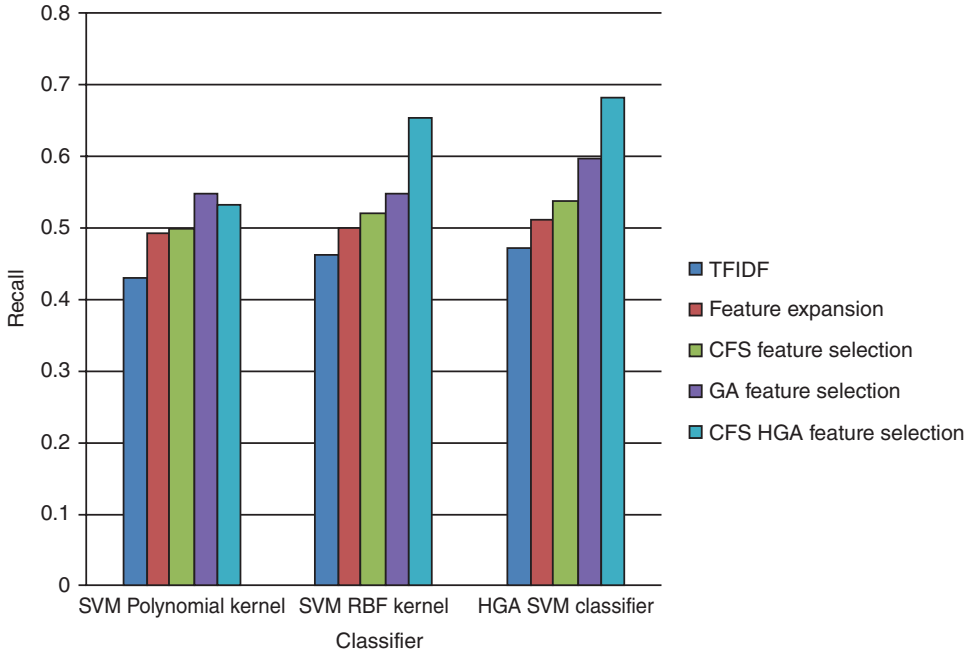

FIGURE 12 | SVM PolyKernal, RBF Kernal, and HGA-SVM versus recall comparison.

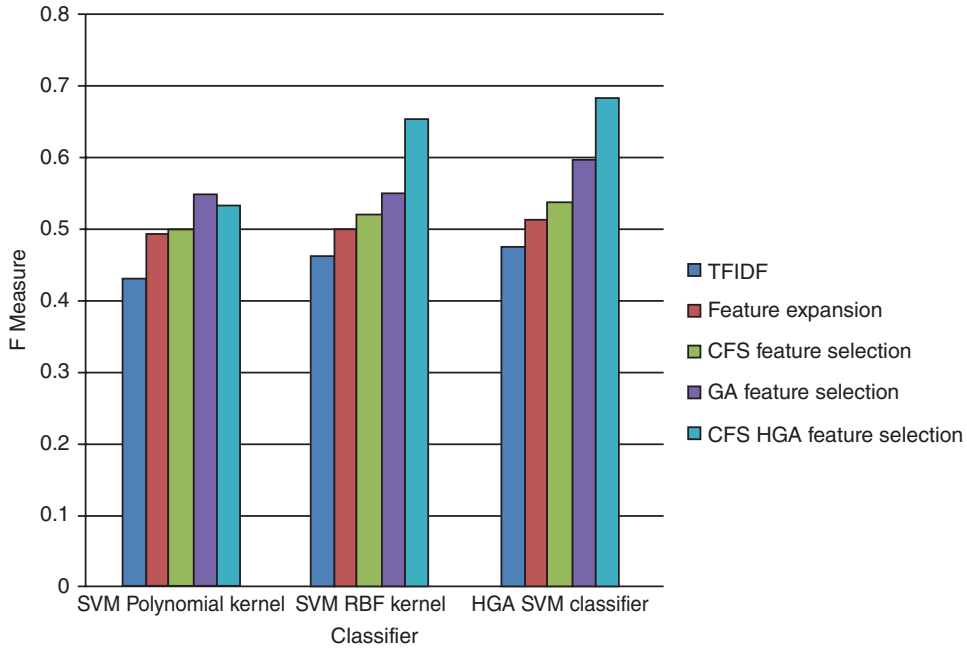

FIGURE 13 | SVM PolyKernal, RBF Kernal, and HGA-SVM versus *F*-measure comparison.

clearly have moved into a real and dominant tool in an assortment of applications, and data mining algorithms have continued to be a significant part of the recommendation process.

A tag classification system is proposed and evaluated using CiteULike dataset. A feature expansion using hypernym and hyponym and CFS-HGA is proposed and compared with other existing feature selection methods, such as TF-DF, CFS, and GA feature selection. Also, a classifier HGA-SVM is proposed to optimize the RBF kernel. It is observed from the experimental results that the precision, recall for the proposed methods CFS-HGA feature selection and HGA-SVM achieves the best performance.

The semantic annotation of web resources is a very expensive, time-consuming process and also

needs extra cost to construct the ontology. Researches so far indicate that constructed ontology cannot fully reflect the individual view of a resource according to the individuality of the users. Therefore, the appropriateness of suggested tags is getting reduced along with the provision of resource descriptions that allow users to 'tap into the long tail' and to find the niches that are relevant for them. Furthermore, future research is required in the semantic annotation of ontologies.

There are a variety of potential directions for future research in the area of Folksonomy-based tag recommendation system.

- One area of future research is to examine the degree to which the retrieval interfaces influence the tag assignment process in catering to the user needs.
- The major advantage of the tag recommendation is its accessibility of large-scale, real-life data from a wide range of tagging systems. Tag recommendations strategies are provided from the perspective of the resources, i.e., the user can train the recommendation system to predict their future tagging decisions.

REFERENCES

1. O'Reilly T. *What Is Web 2.0: Design Patterns And Business Models For The Next Generation Of Software*. O'Reilly Media; 2005.
2. Marlow C, Naaman M, Boyd D, Davis M. Position paper, tagging, taxonomy, Flickr, article, ToRead. In: *Proceedings of the International Workshop on Collaborative Web Tagging Workshop at WWW*, Edinburgh, Scotland, 2006.
3. Strohmaier M, Korner C, Kern R. Why do users tag? Detecting users' motivation for tagging in social tagging systems. In: *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM2010)*, Washington, DC, 2010, 339–342.
4. Strohmaier M. Purpose tagging: capturing user intent to assist goal-oriented social search. In: *Proceedings of the ACM Workshop on Search in Social Media*, Napa Valley; 2008, 35–42.
5. Golder SA, Huberman BA. The structure of collaborative tagging systems. Technical Report, Information Dynamics Lab, HP Labs, 2005.
6. Lipczak M, Milios E. The impact of resource title on tags in collaborative tagging systems. In: *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, New York, NY, 2010b, 179–188.
7. Solskinnsbakk G, Gulla JA. Semantic annotation from social data. *Smart Sens Context* 2009, 5741:66–76.
8. Zubiaga A. Tags vs shelves: from social tagging to social classification. In: *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, New York, NY, 2011, 93–102.
9. Gupta M, Li R, Yin Z, Han J. Survey on social tagging techniques. *ACM SIGKDD Explor Newslett* 2010, 12:58–72.
10. Ames M, Naaman M. Why we tag: motivations for annotation in mobile and online media. In: *Proceedings of the 25th ACM Conference on Human Factors in Computing Systems (CHI'07)*, New York, NY, 2007, 971–980.
11. Golder SA, Huberman BA. Usage patterns of collaborative tagging systems. *J Inf Sci* 2006, 32:198–208.
12. Sen S, Lam SK Rashid AM, Cosley D, Frankowski D, Osterhouse J, Harper MF, Riedl J. Tagging, communities, vocabulary, evolution. In: *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work (CSCW'06)*, New York, NY, 2006, 181–190.
13. Xu Z, Fu Y, Mao J, Su D. Towards the semantic web: collaborative tag suggestions. In: *Proceedings of the International Workshop on Collaborative Web Tagging Workshop at WWW*, Edinburgh, Scotland, 2006.
14. Bischoff K, Firan CS, Nejdl W, Paiu R. Can all tags be used for search? In: *Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*, New York, NY, 2008, 203–212.
15. Cantador I, Castells P, Bellogín A. An enhanced semantic layer for hybrid recommender systems: application to news recommendation. *Int J Semantic Web Inf Syst* 2011, 7:44–78.
16. Aggarwal CC. *Social Network Data Analytics*. New York, NY: Springer; 2011.
17. Lipczak M, Milios E. Learning in efficient tag recommendation. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*, New York, NY, 2010a, 167–174.
18. Pirolli P. Rational analyses of information foraging on the web. *Cognit Sci* 2005, 29:343–373.
19. Cattuto C, Loreto V, Pietronero L. Semiotic dynamics and collaborative tagging. *Proc Natl Acad Sci* 2007, 104:1461–1464.
20. Harry H, Valentin R, Hana S. The complex dynamics of collaborative tagging. In: *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, New York, NY, 2007, 211–220.
21. Frank MJ, Cohen MX, Sanfey AG. Multiple systems in decision making: a neurocomputational perspective. *Curr Dir Psychol Sci* 2009, 18:73–77.

22. Halpin H, Robu V, Shepherd H. The complex dynamics of collaborative tagging. In: *Proceedings of 16th ACM International Conference on World Wide Web*, 2007, 211–220.
23. Rae A, Sigurbjörnsson B, van Zwol R. Improving tag recommendation using social networks. In: *Proceedings of the International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*, New York, NY, 2006, 92–99.
24. Subramaniyaswamy V, Chenthur Pandian S. Effective tag recommendation system based on topic ontology using Wikipedia and WordNet. *Int J Intell Syst* 2012, 27:1034–1048.
25. Jaschke R, Marinho L, Hotho A, Thieme LS, Stumme G. Tag recommendations in folksonomies. In: *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Warsaw, Poland, 2007, 506–514.
26. Lipczak M, Hu Y, Kollet Y, Milios E. Tag sources for recommendation in collaborative tagging systems. In: *Proceedings of the ECML/PKDD Discovery Challenge 2009 Workshop*, Bled, Slovenia, 2009, 497:157–172.
27. Lipczak M, Milios E. Efficient tag recommendation for real life data. *ACM Trans Intell Syst Technol* 2011, 3:2:1–2:21.
28. Melville P, Sindhwani V. Recommender Systems. *Commun ACM* 1997, 40:56–58.
29. Cantador IK, Jose JM. Categorising social tags to improve folksonomy-based recommendations. *Web Semant* 2011, 9:1–15.
30. Pazzani MJ, Billsus D. Content-based recommendation systems. *Lect Notes Comput Sci* 2007, 4321:325–341.
31. Marinho LB, Thieme LS. Collaborative tag recommendations. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, eds. *Data Analysis and Knowledge Organization, Machine Learning and Applications Classification*, vol. 37. Berlin, Heidelberg: Springer; 2005, 533–540.
32. Dattolo A, Ferrara F,Tasso C. The role of tags for recommendation: a survey. In: *Proceedings of the Third International Conference on Human System Interaction*, Rzeszow, Poland, 2010, 548–555.
33. Burke R. Hybrid recommender systems: survey and experiments. *User Model User-Adapt Interact* 2002, 12:331–370.
34. Burke R. Hybrid web recommender systems. In: Brusilovsky P, Kobsa A, Nejdl W, eds. *The Adaptive Web*. Lecture Notes in Computer Science 4321. Berlin: Springer; 2007, 377–408. ISBN: 978-3-540-72078-2.
35. Sigurbjornsson B, van Zwol R. Flickr tag recommendation based on collective knowledge. In: *Proceedings of the 17th ACM International Conference on World Wide Web*, New York, NY, 2008, 327–336.
36. Jiang M, Cui P, Liu R, Yang Q, Wang F, Zhu W, Yang S. Social contextual recommendation. In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12)*, ACM, New York, NY, 2012, 45–54.
37. Mishne G. AutoTag: a collaborative approach to automated tag assignment for weblog posts. In: *Proceedings of the 15th international conference on World Wide Web*, New York, NY, 2006, 953–954.
38. Zhou T, Lü L, Zhang Y. Predicting missing links via local information. *Eur Phys J B* 2009, 71:623–630.
39. Maguitman AG, Cecchini RL, Lorenzetti CM, Menczer F. Using topic ontologies and semantic similarity data to evaluate topical search. In: *Proceedings of the 36th International Latin American Informatics Conference (CLEI)*, 2010.
40. Dix A, Katifori A, Lepouras G, Vassilakis C, Shabir N. Spreading activation over ontology-based resources from personal context to web scale reasoning. *Int J Semant Comput* 2012, 4:59–102.
41. Crestani F. Application of spreading activation techniques in information retrieval. *Artif Intell Rev* 1997, 11:453–482.
42. Hussein T, Neuhaus S. Semantic models for adaptive interactive systems (SEMAIS), In: *1st Workshop in Conjunction with the International Conference on Intelligent User Interfaces (IUI)*, Hong Kong, China, 2010.
43. Subramaniyaswamy V, Vijayakumar V, Indragandhi V. A review of ontology-based tag recommendation approaches. *Int J Intell Syst* 2013, 28:1054–1071.
44. Sieg A, Mobasher B, Burke RD. Learning ontology-based user profiles: a semantic approach to personalized web search. *IEEE Intell Inf Bull* 2007, 8:7–18.
45. Durao F, Dolog P. Extending a hybrid tag-based recommender system with personalization. In: *Proceedings of the 2010 ACM Symposium on Applied Computing (SAC)*, Sierre, Switzerland, 2010, 1723–1727.
46. Subramaniyaswamy V, Chenthur Pandian S. Topic ontology-based efficient tag recommendation approach for blogs. *Int J Comput Sci Eng* 2014, 9:177–187.
47. Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data. In: Maimon O, Rokach L, eds. *Data Mining and Knowledge Discovery Handbook*. 2nd ed. Heidelberg: Springer; 2010, 667–685. ISBN: 978-0-387-09822-7.
48. Huang W, Kataria S, Caragea C, Mitra P, Giles CL, Rokach L. Recommending citations: translating papers into references. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*, New York, NY, 2012, 1910–1914.
49. Zhou T, Ma H, Lyu M, King I. Userrec: a user recommendation framework in social tagging systems. In: *Proceedings of the 24th AAAI Conference*, Atlanta, GA, 2010, 1486–1491.
50. Liu Z, Huang W, Zheng Y, Sun M. Automatic keyphrase extraction via topic decomposition. In:

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'10), Massachusetts, MA, 2010, 366–376.

51. Yagnik S, Thakkar P, Kotecha K. Recommending tags for new resources in social bookmarking system. *Int J Data Min Knowl Manage Process* 2014, 4:19–32.
52. Tuarob S, Pouchard LC, Giles LC. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In: *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'13)*, New York, NY, 2013, 239–248.
53. Subramaniyaswamy V. Automatic topic ontology construction using semantic relations from WordNet and Wikipedia. *Int J Intell Inf Technol* 2013, 9:61–89.
54. Farooq, U., Kannampallil, T. G., Song, Y., Ganoe, C. H., Carroll, J. M., & Giles, L. Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In: *Proceedings of the 2007 international ACM conference on supporting group work*, New York, 2007, 351–360.
55. Li R, Guo X. Improved term selection algorithm based on variance in text categorization. In: *Proceedings of the 2nd International Conference on Systems Engineering and Modeling (ICSEM-13)*, Paris, France, 2013.
56. Aggarwal CC, Zhai C. *Mining Text Data*. Springer Science Business Media; 2012.
57. Cann R. Sense relations. In: Maienborn C, Von Heusinger K, Portner P, eds. *Semantics: An International Handbook of Natural Language Meaning*, vol. 1. Berlin: Mouton de Gruyter; 2011, 456–479.
58. Roy D, Sarkar S, Ghose S. Automatic extraction of pedagogic metadata from learning content. *Int J Artif Intell Educ* 2008, 18:97–118.
59. Karegowda AG, Manjunath AS, Jayaram MA. Comparative study of attribute selection using gain ratio and correlation based feature selection. *Int J Inf Technol Knowl Manage* 2010, 2:271–277.
60. Hall, MA. Correlation-based feature selection for machine learning. PhD Dissertation, The University of Waikato, 1999.
61. Goldberg DE, Holland JH. Genetic algorithms and machine learning. *Mach Learn* 1988, 3:95–99.
62. Karuppathal R, Palanisamy V. Hybrid GA-SVM for feature selection to improve automatic Bayesian classification of brain MRI slice. *Life Sci J* 2013, 10:2273–2280.
63. Tripathy RK, Acharya A, Choudhary SK. Gender classification from ECG signal analysis using least square support vector machine. *Am J Signal Process* 2012, 2:145–149.
64. Kaur S, Sandhu ERK. A survey on enhanced human identification using gait recognition based on neural network and support vector machine. *Int J Appl Innov Eng Manage* 2013, 2:24–27.
65. Baldoni M, Baroglio C, Patti V, Rena P. From tags to emotions: ontology driven sentiment analysis in the social semantic web. *Intelligenza Artificiale* 2012, 6:41–54.
66. Breese J, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, Madison, WI, 1998, 43–52.
67. Brooks CH, Montanez N. Improved annotation of the blogopshere via autotagging and hierarchical clustering. In: *Proceedings of the 15th International Conference on World Wide Web*, New York, NY, 2006.
68. Chi Y, Zhu S, Song X, Tatemura J, Tseng BL. Structural and temporal analysis of the blogosphere through community factorization. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, 2007, 163–172.
69. Gemmell J, Schimoler T, Mobasher B, Burke R. Resource recommendation in social annotation systems: a linear-weighted hybrid approach. *J Comput Syst Sci* 2012, 78:1160–1174.
70. Géry M, Haddad H. Evaluation of web usage mining approaches for user's next request prediction. In: *Fifth International Workshop on Web Information and Data Management*, Boston, MA, 2003, 74–81.
71. Geyer-Schulz A, Hahsler M. Evaluation of recommender algorithms for an internet information broker based on simple association rules and on the repeat-buying theory. In: *Fourth WEBKDD Workshop: Web Mining for Usage Patterns & User Profiles*, Edmonton, 2002, 100–114.
72. Good N, Schafe, JB, Konstan JA, Borchers A, Sarwar B, Herlocker J, Riedl J. Combining collaborative filtering with personal agents for better recommendations. In: *Proceedings of Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, Menlo Park, CA, 1999, 439–446.
73. Herring SC, Scheidt LA, Bonus S, Wright E. Bridging the gap: a genre analysis of weblogs. In: *Proceedings of the 37th Hawaii International Conference on System Sciences*, Washington, DC, 2004.
74. Jin X, Lin CX, Luo J, Han J. SocialSpamGuard: a data mining based spam detection system for social media networks. *Proceedings VLDB Endowment* 2011, 4:1458–1461.
75. Kao YH, Chen TS, Lee WB, Chen RC, Huang CC, Lin MC, Wang YL. Key training items search of manufacturing assessment based on TTQS and GA-SVM. *Inf Technol J* 2013, 12:756–762.
76. Krause B, Schmitz C, Hotho A, Stumme G. The anti social tagger detecting spam in social bookmarking systems. In: *Proceedings of the Fourth International ACM Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, New York, NY, 2008, 61–68.
77. Krötzsch M, Vrandecic D, Volkel M, Haller H, Studer R. Semantic wikipedia. *J Web Semant* 2007, 5:251–261.

78. Liu B, Zhai E, Sun H, Chen Y, Chen Z. Filtering spam in social tagging system with dynamic behavior analysis. In: *Proceedings of International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, Athens, Greece, 2009, 95–100.

79. Melville P, Gryc W, Lawrence RD. Sentiment analysis of blogs by combining lexical knowledge with text classification. In: *Proceedings of Fifteenth ACM SIGKDD International Conference on knowledge Discovery and Data Mining*, New York, NY, 2009, 1275–1284.

80. Pang B, Lee L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2008, 2:1–135.

81. Park Y, Heo GM, Lee R. Blogging for informal learning: analyzing bloggers' perceptions using learning perspective. *Educ Technol Soc* 2011, 14:149–160.

82. Sarwar B, Karypis G, Konstan JA, Reidl J. Item-based collaborative filtering recommendation algorithms. In: *Proceedings of the Tenth International Conference on World Wide Web*, New York, NY, 2001, 285–295.

83. Schafer JB, Konsta JA, Riedl J. E-Commerce recommender applications. *Data Min Knowl Discov* 2001, 5:115–152.

84. Snasel V, Moravec P, Pokorny J. WordNet ontology based model for web retrieval. In: *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)*, Washington, DC, 2005, 220–225.

85. Wolf J, Aggarwal C, Wu K-L, Yu P. Horting hatches an egg: a new graph-theoretic approach to collaborative filtering. In: *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, *San Diego*, *CA*, 1999.

86. Zhuang J, Hoi SCH, Sun A. On profiling blogs with representative entries. In: *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, New York, NY, 2008, 55–62.

