International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11
ISSN: 2454-4248
396 – 402

Collaborative Tagging and Taxonomy by Vector Space Approach

[1]R. Sathya,
M. Phil Scholar, Department of Computer Science,
Selvamm Arts and Science College (Autonomous) Namakkal (Tk) (Dt) – 637003.

[2]Mrs. K.V. Sumathi,
MCA., M. Phil, Assistant Professor, Department of Computer Science,
Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003.

[3] Mrs. K. K. Kavitha,
M.C.A., M.Phil., SET., (Ph.D)., Vice Principal, Head of the Department of Computer Science, Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003

Abstract:- Collaborative tagging or group tagging is tagging performed by a group of users usually to support in re-finding the items. The limberness of tagging allows users to classify their collections of items in the ways that they find useful, but the personalized variety of expressions can present challenges when searching and browsing. When users can liberally choose tags (users create and apply public tags to online items as different to selecting terms from a proscribed terminology based on the users feedback), the resulting metadata can consist of homonyms (the same tags used with dissimilar implication) and synonyms (multiple tags for the same concept) which may direct to inappropriate connections between items and wasteful searches for information about a subject.

Collaborative tagging requires the enforcement of method that enables users to protect their privacy by allowing them to hide certain user-generated contents without making them useless for the purposes they have been provided in a given online service. This means that privacy-preserving mechanisms must not harmfully affect the service truthfulness and usefulness.The proposed approach defends the user privacy to a certain level by reducing the tags that make a user profile let somebody see partiality toward certain categories of interest or feedback.

*****

INTRODUCTION

GENERAL BACKGROUND

Data mining also popularly referred to as Knowledge Discovery from Data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or catchable in large databases, data warehouses, the Web, other massive information repositories, or data streams.

Data mining is a multidisciplinary field, draws work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. The discoveries of patterns hidden in large data sets are focusing on issues relating to their feasibility, usefulness, effectiveness, and scalability.

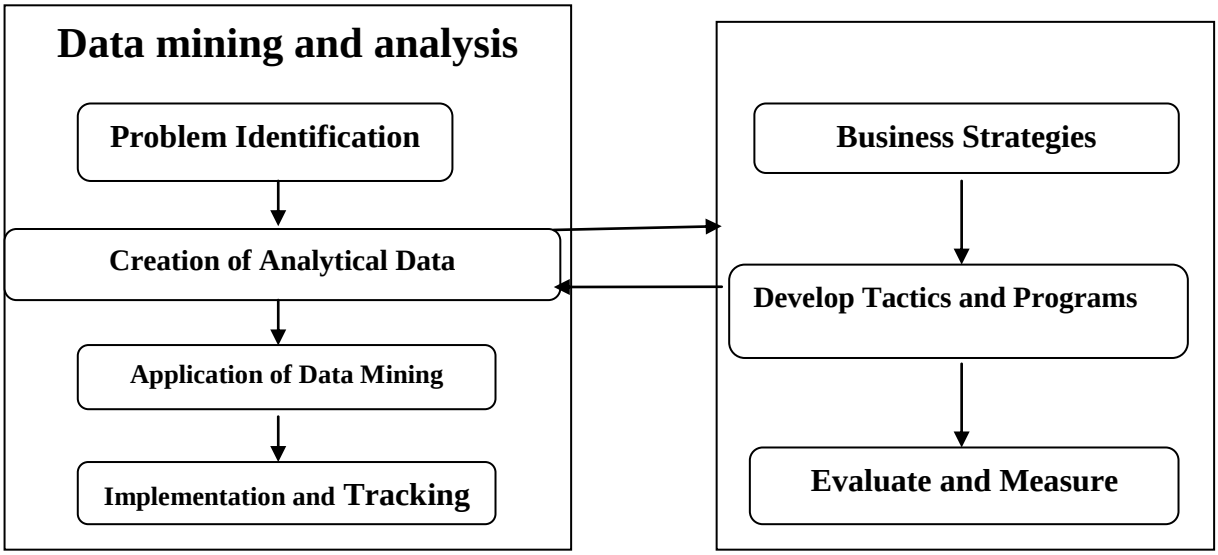


Fig 1.1 Data Mining Basics

 ISSN: 2454-4248
396 – 402

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. It is useful for computing science students, application developers, and business professionals, as well as researchers.

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.

natural selection in a design based on the concepts of natural evolution.

DIFFERENT LEVELS OF DATA MINING

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and

Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

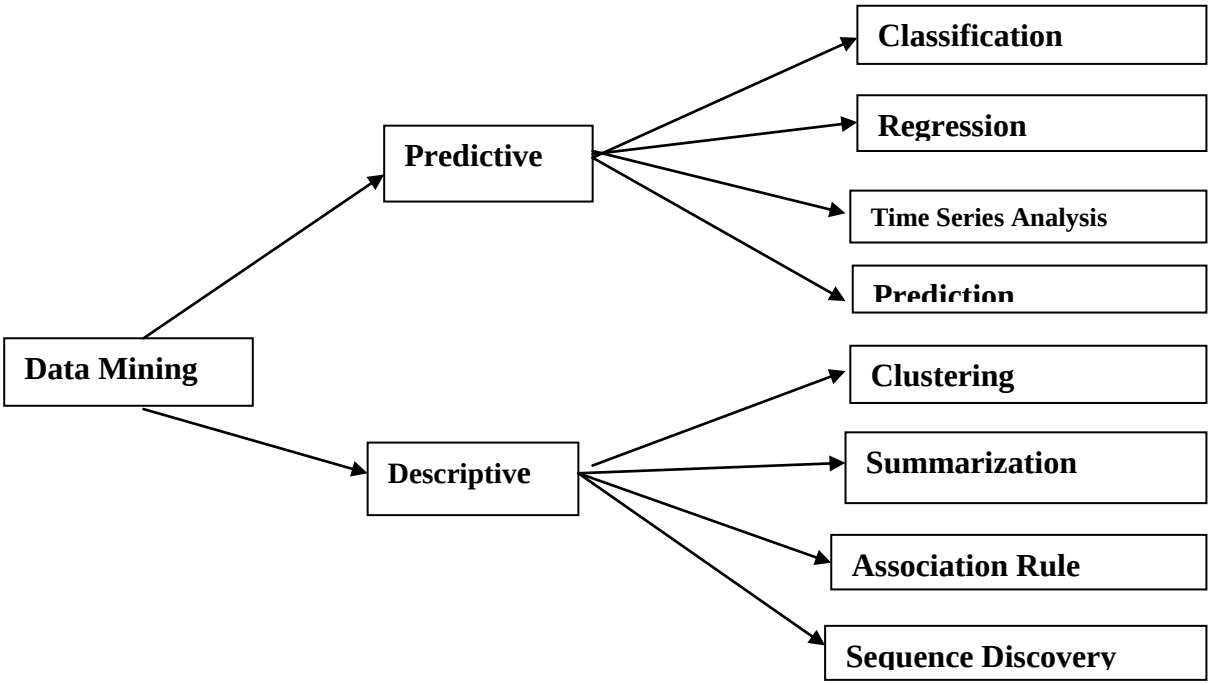


Fig 1.2 Data Mining Level

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

PURPOSE OF DATA MINING

The important general difference in the focus and purpose between Data Mining and the traditional Exploratory Data Analysis (EDA) is that Data Mining is more oriented towards applications than the basic nature of the underlying phenomena.

 ISSN: 2454-4248
396 – 402

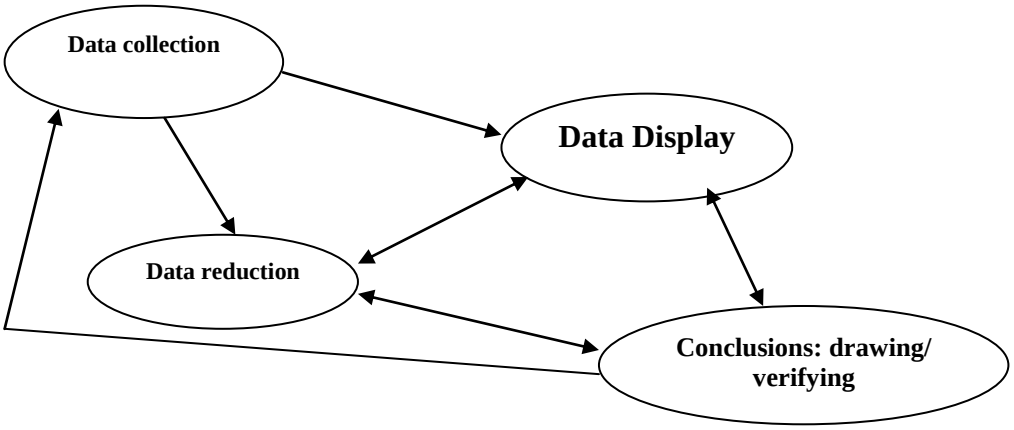


Fig 1.3 Data Mining Usages

Data Mining is often considered to be "a blend of statistics, AI [artificial intelligence], and data base research which until very recently was not commonly recognized as a field of interest for statisticians, and was even considered by some a dirty word in Statistics.

WEB MINING

Web mining helps to extract useful information from the web pages. Various we mining techniques are used to extract knowledge from the web data, web documents and hyperlinks between the documents. Where the web is universal information platform space which can be accessed by companies, universities, businessman etc. Generally, web hold there are numerous sources of information like internal sources and external sources.

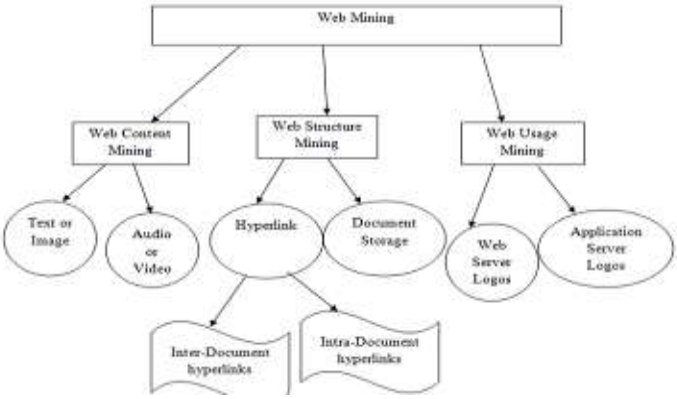


Fig 1.4 Web Mining

Internal sources are those which include personal information of any organization and external sources are those which include information of clients, vendors, suppliers, intranet and extranet etc. In this research paper, we can divide us mining into three categories which are listed as:

a) Web Structure Mining.
b) Web Usage Mining.
c) Web Content Mining.

Web Structure Mining

It consists of web pages as nodes as hyperlinks and edges connecting related pages. It basically tells the structural layout of the web. It also used the connectivity among websites that are called “Hyperlinks”. Hyperlinks are further divided into two categories which are listed as below:-

- Internal hyperlinks that lead to pages within the same web page.
- External hyperlinks that lead to other web pages.
- Document structure is basically a schema language for XML which helps to describing a valid XML documents.

Web Usage Mining

It holds the knowledge discovered by users which are navigating through the websites. files. It is further divided into two categories which are listed as follows

 ISSN: 2454-4248
396 – 402

- **Application Server Data**: It holds the business transactions and also makes their repository in applications server log.
- **Web Server Data:** In these logs are made by the web server. It also includes the field of IP address means the number of web pages accessed with access times.

Web Content Mining

It holds the knowledge discovery by going through the web pages contents like image, videos etc. Intelligent agents help to solve the problem of indexing in search engines otherwise it will result in delivery imprecise results but information overloading. It also helps to select much more relevant documents.

WEB MINING SECURITY

Web mining has emerged as an important branch of data mining. This is mainly due to the tremendous amount of information available from the Web, which attracted many research communities, and the recent interest of e-commerce.

OBJECTIVES

- To introduce a user-assisted friend grouping mechanism that enhances traditional group-based policy management approaches. Minister to friend grouping leverages proven clustering techniques to aid users in grouping their friends more effectively and efficiently.
- To found measurable agreement between clusters and user-defined relationship groups. In addition, user perceptions of the improvements should be encouraging.
- To introduce a new privacy management model that is an enhancement over conventional group-based policy management approaches.
- To leverage a user's memory and attitude of their friends to situate policies for other associated friends, which refer to as Same-As Policy Management.

PROBLEM DEFINITION

Collaborative tagging systems such as Delicious, last.FM, and Bibsonomy are valuable components of the Social Web. They allow users, firstly, to organize their own data with a level of freedom not possible in traditional taxonomic filing systems whether it is web bookmarks, music collections, or academic journal references. Secondly, they provide users a means to openly share this information so that friends and colleagues can easily communicate with each other about their latest discoveries. The major issues of the existing approaches are

- To allow anyone to utilize the collective knowledge of others for discovering new resources and perhaps even new friends. These benefits, however, are only as powerful as the system is trustworthy. As with any open adaptive system, maintaining the integrity of a tagging system presents a considerable logistical problem.
- It is loosely classify resources based on end-user's feedback, expressed in the form of free-text labels (i.e., tags). The novelty of such an approach to content/resource categorization has been seen, in recent years, as a challenging research topic.
- The undefined semantics of tags, which are per se ambiguous and expressed in multiple languages, makes it difficult to enforce semantic interoperability and to grant a reasonable level of accuracy when determining the "meaning" of a tag.
- Tag prediction concerns the possibility of identifying the most probable tags to be associated with a non-tagged resource, whereas tag recommendation is meant to suggest to users the tags to be used to describe resources they are book marking. In both cases, existing approaches apply techniques usually enforced in recommendation systems.
- Another interesting issue concerns the exploitation of the "explicit" relationships between users (i.e., the actual social network underlying a folksonomy to address issues like annotation relevance and/or trustworthiness. Privacy protection in social tagging services is another issue that has not been thoroughly investigated.

SCOPE OF RESEARCH WORK

This proposed system's methodology could be implemented in the following real world applications:

- Marketing strategy analysis
- Social communication application of a individual organization
- Rating process over online

An enhanced collaborative tagging system that consists of a "traditional" book marking service, such as Delicious, and two main additional services built on top tree view. Such services address two main issues. The former allows end users to specify policies that can be used either to explicitly denote resources of interests or to enforce blocking conditions on the browsed data. The latter features a specific PET, namely, tag suppression, to preserve the privacy of registered users by hiding the specific characteristics of their profiles. Such architecture is a specific implementation of the multilayer framework presented. with the relevant difference that in [5] the privacy layer is missing. Lastly, we would also like to emphasize that our approach is not limited to the specific book marking application here contemplated, i.e., Delicious. As a matter of fact, it could be built on top of any collaborative tagging system.

Nevertheless, if tags were not sensible information per se, they could easily be exploited to infer users' personal

 ISSN: 2454-4248
396 – 402

information, such as personal interests, preferences, and opinions. This is even easier when it is possible to statistically analyze huge collections of tags as those made publicly available by social bookmarking services, thus obtaining accurate tag-based user profiles. In this field, privacy-preserving techniques should guarantee, at the same time privacy protection and the correctness of the results obtained by analyzing the data set.

LITERATURE REVIEW

Wu, L. Zhang, and Y. Yu et al[2] proposed explore a complement approach that focuses on the"social annotations of the web" which are annotations manually made by normal web users without a pre-defined formal ontology. Compared to the formal annotations, although social annotations are coarse-grained, informal and vague, they are also more accessible to more people and better reflect the web resources' meaning from the users' point of views during their actual usage of the web resources. But this system focuses only on bookmarking web based system. Also this system not considers the tagging approach.

B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho,and S. Gerd et al[3] proposed Here they build an evaluation framework to compare various general folksonomy-based similarity measures, which are derived from several established information theoretic, statistical, and practical measures. Their framework deals generally and symmetrically with users,tags, and resources. For evaluation purposes we focus onsimilarity between tags and between resources and consider different methods to aggregate annotations across users.This approach shows that they can define relations of users from tags. This is important from privacy preserving point of view.

C. Marlow, M. Naaman et al[4] provide a short description of the academic related work to date. They offer a model of tagging systems, specifically in the context of web-based systems, to help us illustrate the possible benefits of these tools. Since many such systems already exist, they provide a taxonomy of tagging systems to help inform their analysis and design, and thus enable researchers to frame and compare evidence for the sustainability of such systems. They also provide a simple taxonomy of incentives and contribution models to inform potential evaluative frameworks. They present a preliminary study of the photo-sharing and tagging system Flicker to demonstrate our model and explore some of the issues in one sample system. Hence this paper is just giving us the basic idea about how tag functionality works in web based systems. This is important to clear basic ideas about tagging.

METHODOLOGY

INTRODUCTIONTAG

A tag is a user-contributed metadata, providing a mean of information or content item, created freely by users with personally salient keywords or labels, known as tags. The process of labeling is called tagging. Users can re-find the information later by means of those tags that they have created. Also, by tagging users can store resources for their future retrieval.

COLLABORATIVE TAGGING SYSTEM

Collaborative tagging is a classification by the users and for the users. It is a social, decentralized and complex network where many annotations, generally provided by interrelated groups of individuals, are organized so to link resources and tags. Each resource item can be associated with many different tags, rather than with a single branch of a hierarchy.

With tags chosen freely from common language and associated with web resources that are interesting for users (such as photographs, videos, web links and documents), collaborative tagging offers a sense of community in managing resources and results in a process of knowledge construction. Users can share their resources with others, discover resources through the collaborative network, and contact people with similar interests. The benefit of collaborative tagging systems comes from the many views of the mass, rather than from a dominant opinion supplied by a few.

In fact, the form of tagging tends to stabilize over time because people usually choose to use the tags in three ways:

- Imitation, users are easily affected by the tags that were previously applied by others to the same page;
- Habit, users re-use tags that they have already used on other pages respect to their background and culture;
- Recommendation, users choose tags that are suggested by a given interface.

COLLABORATIVE METHODS

Several recommendation systems use a hybrid approach by combining collaborative and content-based methods, which helps to avoid certain limitations of content-based and collaborative systems. Different ways to combine collaborative and content-based methods into a hybrid recommender system can be classified as follows:

- Implementing collaborative and content-based methods separately and combining their predictions,
- Incorporating some content-based characteristics into a collaborative approach,
- Incorporating some collaborative characteristics into a content-based approach, and
- Constructing a general unifying model that incorporates both content-based and collaborative characteristics.

TAG SUPPRESSION

 ISSN: 2454-4248
396 – 402

In our scenario of collaborative tagging, users tag resources on the web, for example, music, pictures, videos or bookmarks, according to their personal preferences. Users therefore contribute to describe and classify those resources, but this is inevitably at the expense of revealing their profile. To avoid being accurately profiled by tagging systems or in general by any attacker able to collect such information, users may adopt a privacy-enhancing technology based on data perturbation. The data-perturbative technology considered in this work is tag suppression, a technique that allows a user to refrain from tagging certain resources in such a manner that the profile resulting from this perturbation does not capture their interests so precisely.

PROPOSED METHOD PROCEDURE

The proposed methodology of the thesis is implemented with the following modules and procedures to experiment the proposed methodology.

- User Registration
- Post Content
- View Posts
- General Suppression Word
- Strict Suppression Word
- My Tag Cloud
- Add Child User
- View Child Users
- Policy/ Resource Recommendation
- Policy/ Parental Control
- Change Password

EXPERIMENTAL RESULTS AND DISCUSSION

IMPLEMENTATION SOFTWARE

The empirical system is designed and implemented by using the Microsoft visual studio .net as a front end tool. And the coding language used is C#.net. Microsoft SQL Server used as a back end tool. Visual studio is an integrated development environment which is used in this thesis for designing the thesis experiments.

FEATURES OF C#.NET

C# is a Microsoft‘s new language designed for its new platform ".NET". It is fully object oriented language like java and is the first component-oriented language. Because it contains integral supports for writing the software components. C# is designed for building robust, reliable and durable components to handle real world application. The C# language specification stated the objectives and features of C#:

- It is simple, modern, general purpose and object oriented programming language.
- This provides a support for the software analysis principles such as strong type checking, array bounds checking, detection of attempts to use uninitialized variables and automatic garbage collection.
- It is useful for developing software components which are suitable for deployment in the distributed environments. This supports internationalization.

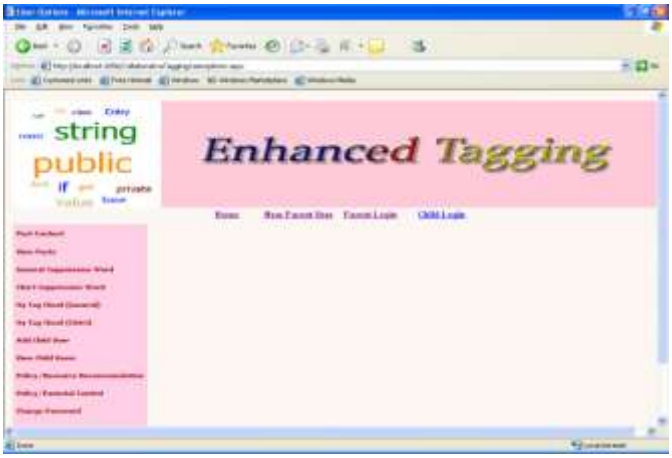
CHARACTERISTICS OF C#

- **Garbage Collection:** the memory management feature leads all managed objects. Garbage collection is a feature .NET. The C# uses it during the runtime.
- **Indexes:** C# has indexes which help to access value in a class with an array like syntax programs.
- **Exception Handling:** .NET standardizes the exception handling across languages. C# offers the conditional keyword to control the flow and make the code more readable.
- **Versioning:** C# programming supports this versioning. The .NET solves the versioning problem and enables the software developer to specify version dependencies between the different pieces of software.
- **Extensive Inter-operability:** All enterprise software application can be managed easily by type safe environment. This extensive inter-operability makes C# which is the obvious choice for the software developers.

EXPERIMENTAL RESULTS

INPUT FORM DESIGN USER LOGIN PAGE

USER OPTION



 ISSN: 2454-4248
396 – 402

CONCLUSION

Collaborative tagging is currently an enormously popular online service. Although nowadays it is basically used to support resource search and browsing, its achievable is still to be demoralized. One of these potential applications is the provision of web access functionalities such as content filtering and innovation. For this to become a reality, however, it would be necessary to extend the architecture of current collaborative tagging services so as to include a policy layer that supports the enforcement of user inclinations.

Collaborative tagging has been gaining popularity, it have been become more obvious the need for privacy safeguard; not only because tags are susceptible information but also because of the risk of cross referencing. In addition to the existing system approaches, the proposed system takes care of multi language tagging.

FUTURE ENHANCEMENTS

A privacy preserving collaborative tagging if functional to content with various languages, and then it becomes more efficient to beneficial to end users. Future work includes the development of a full prototype for the experimented system and it's testing and use in further scenarios.

The proposed methodology can be enhanced and implemented with the various applications like

- Official information sites,
- Employees skills registry portals,
- Other official or personal related confidential information management system through online.
- In future the algorithm will be modified or redefined to improve the efficiency of the methodology.

REFERENCES

[1] Adomavicius.G and Tuzhilin.A, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. Knowledge Data Eng., vol. 17,no. 6, pp. 734-749, June 2005.

[2] Barnes.S.B, "A Privacy Paradox: Social Networking in the United States," First Monday, vol. 11, no. 9, Sept. 2006.

[3] Bischoff .K ,Firan . C . S, Nejdl . W, and Paiu . R , "Can All Tags Be Used for Search?" Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), pp. 193-202, 2008.

[4] Bundschus.M, Yu.S, Tresp.V, Rettinger.A , Dejori.M , and Kriegel.H.P, "Hierarchical Bayesian Models for Collaborative Tagging Systems," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 728-733,2009.

[5] Carminati.B, Ferrari.E , and Perego.A, "Combining Social Networks and Semantic Web Technologies for Personalizing Web Access," Proc. Fourth Int'l Conf. Collaborative Computing: Networking, Applications and Work sharing, pp. 126-144, 2008.

[6] Frı´as-Martinez.E, Cebria´n.M , and Jaimes.A , "A Study on the Granularity of User Modeling for Tag Prediction," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence Intelligent Agent Technology (WIIAT), pp. 828-831, 2008.