P2PDocTagger: Content management through automated P2P collaborative tagging

Hock Hee Ang Vivekanand Gopalkrishnan Wee Keong Ng Steven C.H. Hoi
Nanyang Technological University, Singapore

ABSTRACT

As the amount of user generated content grows, personal information management has become a challenging problem. Several information management approaches, such as desktop search, document organization and (collaborative) document tagging have been proposed to address this, however they are either inappropriate or inefficient. Automated collaborative document tagging approaches mitigate the problems of manual tagging, but they are usually based on centralized settings which are plagued by problems such as scalability, privacy, etc. To resolve these issues, we present P2PDocTagger, an automated and distributed document tagging system based on classification in P2P networks. P2PDocTagger minimizes the efforts of individual peers and reduces computation and communication cost while providing high tagging accuracy, and eases of document organization/retrieval. In addition, we provide a realistic and flexible simulation toolkit – P2PDMT, to facilitate the development and testing of P2P data mining algorithms.

1. INTRODUCTION

The amount of personal data, such as emails, documents, photos, videos, etc., has been rapidly growing, calling for efficient personal information management (PIM) systems. While desktop search applications such as Windows Search[1], Google Desktop[2], etc., provide an easy way to locate documents, they need appropriate search phrases in order to be useful. Hence, browsing the file system is still the preferred way for document retrieval.

To browse in an efficient manner, documents need to be organized properly. However, with a large amount of content, manual organization becomes a very tedious task. Automatic methods that are based on supervised learning paradigms aim to alleviate such problems, but these approaches require a significant amount of labeled data to accurately organize the documents. To address the above issues, Siersdorfer and Sizov [6] proposed a collaborative learning approach that automatically organizes and categorizes documents in a distributed setting. Their approach consolidates the knowledge derived from labeled data of all users to accurately categorize documents.

While document categorization approaches ease the retrieval of files, they have their own limitations. Typically, given a set of predefined categories, these approaches associate each document to a single category, assuming there is no overlap with other categories. However, in reality most documents belong to multiple categories. As a result, tagging based systems have been proposed to model such behaviour, allowing each document to associate with multiple keywords/tags. An important point to note here is that tags are user specified, i.e., they are *open vocabulary* and *non-hierarchical*, unlike categories. In addition, we emphasize here that tags may not necessarily be contained within the documents, or even be associated with the terms in the documents. Therefore, tags cannot be generated by indexing the words/terms of the documents and are usually provided manually by the users, chosen from words which they *feel* are related to, or best describe the document. Examples of such systems include PHLAT [4], which supports manual tagging of documents and allows retrieval by filtering and searching. While tagging provides a more natural way of associating files, the manual process is usually very tedious and time consuming.

In recent years, social tagging or collaborative tagging has acquired significant interest, and has been shown to facilitate the process of finding documents. Moreover, tags generated by different users provide alternate views on the documents, allowing users to stumble upon new information when revisiting a certain topic. However, the issue of manual tagging still plagues such systems, calling for automated tagging approaches [5]. Such approaches typically use data mining or statistical methods to model the assignment of tags. However, they are *centralized* solutions, causing several issues. For instance, when having to deal with a large amount of data from an enormous number of users, *scalability* can become an issue. In addition, system failures can result in catastrophic outcomes. Moreover, centralization of personal data increases the chances of *privacy* leaks and *security* breaches, which are critical when dealing with personal information.

1.1 Novelty and Contributions

To alleviate the issues of categorization and centralized collaborative tagging, we present an *automated* and *dis-*

[1] http://www.microsoft.com/windows/products/winfamily/desktopsearch/default.mspx

[2] http://desktop.google.com/

Proceedings of the VLDB Endowment, Vol. 3, No. 2
Copyright 2010 VLDB Endowment 2150-8097/10/09... $ 10.00.

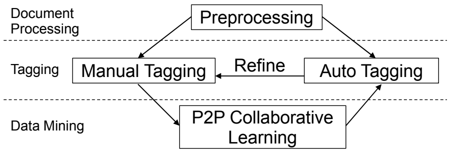


Figure 1: System architecture of P2PDocTagger.

tributed collaborative document tagging system — P2PDocTagger. P2PDocTagger models (based on P2P classification) how documents are tagged in a distributed manner and automatically tags other documents.

Contrary to existing automated collaborative tagging systems, which are typically deployed in the centralized setting, P2PDocTagger functions in a *distributed* manner, allowing physically distributed data to be handled more *efficiently* (without data centralization), and possibly in a privacy preserving manner. In addition, P2PDocTagger *scales well* even in the presence of large amount of data or large number of peers. Moreover, peers are autonomous and hence there is *no single point* of failure in the system. To the best of our knowledge, this is the first automated and distributed collaborative tagging system for PIM.

To realistically test a P2P application, large amount of resources (hundreds or thousands of machines) are required, which are extremely expensive and infeasible for most to acquire. Hence, we developed a flexible P2P data mining simulation toolkit (P2PDMT), as an alternative to facilitate the development and testing of P2P data mining algorithms under *realistic* P2P environments, which has also been used for P2PDocTagger. This paper demonstrates advantages of P2PDocTagger, which is based on state of the art P2P classification framework.

2. SYSTEM OVERVIEW

Figure 1 shows the main components of our system. At this stage, automated tagging is applied only on text documents, but it can be extended to other types. The process of automatic tagging is as follows. First, users select documents (or folders containing documents) that they wish to tag. This ensures that all files processed by the system are approved by the users. Next, these documents are preprocessed for (automated) tagging. In the beginning, when there are no tagged documents in the entire network, users have to manually tag *some* of their documents. Next, the system constructs a global classification model in a distributed manner to learn how users/peers tag their documents. Thereafter, P2PDocTagger can tag documents automatically using the global classification model. While different users may have slightly different opinions on how documents should be tagged, P2PDocTagger achieves localized conflict resolution by allowing users to refine the documents' tag assignments, and this information is used to improve the local and global classification models. Once tags are assigned, they are saved as the files' meta-data, which are supported by numerous operating systems such as GNU/Linux, Mac OS X, Microsoft Windows, etc. In addition to P2PDocTagger, other PIM systems can access these tags for file organization/retrieval. Additional details of each step are discussed next.

Document preprocessing. The document preprocessing in P2PDocTagger is similar to that of information retrieval systems. First, stop words that contain little recognition values (e.g., a, for, and, not, etc), as well as user-specified sensitive words are filtered out from all documents. Next, words are normalized using the porter stemming algorithm to remove the commoner morphological and inflexional endings (English). To allow supervised learning, documents have to be represented as multidimensional feature vectors, i.e., each unique word in the document is represented by a numerical value indicating the weight of the word, in a unique position of the multidimensional feature vector. Or in other words, the attribute id represents the word id and the value of the attributes represents the word frequency in the documents, e.g., a document d is represented by a vector $\{w_1, \ldots, w_m\}^T$, where w_j is the weight of the word represented by id j, and m is the total number of words in the lexicon.

Depending on the P2P collaborative learning algorithm, these document vectors may be shared among peers. Since words are represented by their ids without any sequence information, information about the documents that are revealed to other peers are reduced. In addition, document vectors are only associated with tags, minimizing the revelation of the relationships between the users, documents and tags. As such, no confidential information is revealed, protecting the security and privacy of users.

Automated P2P collaborative tagging. In this work, automated tagging is posed as classification problem. Formally, we want to learn a function (classification model) $f : X \rightarrow Y$, which maps a m-dimensional document vector $d_i = \{w_{i1}, \ldots, w_{im}\}^T \in X$ to a set of corresponding tags $\mathbf{y}_i = \{y_i1, \ldots, y_ih\} \in Y$, where Y is the universe of all tags. In order to construct a classification model that can accurately tag the documents, a large amount of training data, i.e., tagged document examples, are required. This requires substantial effort in manual tagging of the documents, which is too time consuming for a single user. Hence, to reduce these efforts of individual users while generating a large amount of tagged documents for constructing accurate classification models, we propose to exploit P2P networks.

P2P networks contain a large number of peers $p_1, \ldots, p_N$, where N is the total number of peers, who share their individual resources, such as storage space, bandwidth, CPU cycles, etc. The aim of using P2P networks is to consolidate knowledge of the small amount of tagged documents of the large number of individual peers, which is equivalent to that of a large training dataset. In addition, using classification algorithms designed for the P2P networks can reduce the computation and communication cost of learning from the large amount of distributed data. Based on the P2P classification paradigm, P2PDocTagger learns how documents are tagged by users and shares this knowledge among all peers in the P2P network to automatically tag other documents (automated P2P collaborative tagging). Hence, instead of learning from a single training data set $D = \{d_1, \ldots, d_l\}$, where l is the number of training data, as in the centralized setting, the P2P classification algorithm learns from the training data $D_1, \ldots, D_N$ of all peers. As a result, each peer only needs to contribute a small number of tagged documents (together a large pool of tagged documents is generated) for P2PDocTagger to achieve high tagging accuracy.

The problem of document tagging is a *multi-label* classification problem where each document can be associated with *multiple* tags, unlike *single-label* classification, which associates one document to only *one* tag. However, existing P2P classification approaches [2, 1] are based on the single label classification problem. Hence, in this preliminary work, we simplify the multi-label classification problem into numerous single-label classification problems, so that the P2P classification approaches can be used. Instead of learning a function $f : X \to Y$, where Y is the universe of all tags, for each $c \in Y$, we learn a function $f_c : X \to Y_c$, where the output $y_c \in Y_c = \{0, 1\}$ indicates whether or not the tag is assigned to the document. Here, the binary classifiers are constructed using the one-against-all method, where data from a target tag belongs to one class and all data from other tags belong to another class. As this approach is essentially the same as how SVM (the base classifier of our chosen P2P classification approaches) handles multi-class classification problems, it does not incur additional cost compared with the single label classification approach.

At this point, we want to emphasize that the P2P classification algorithm in P2PDocTagger is a pluggable component. Therefore, if we deploy a privacy preserving P2P classification algorithm, P2PDocTagger will then inherit the privacy preserving property. While there are many choices of P2P classification approaches, we implemented our system using two different approaches, viz., CEMPaR [2] and Pace [1]. CEMPaR and Pace have been shown to achieve classification accuracy comparable to centralized approaches and better than other state of the art P2P classification approaches at lower communication cost. In addition, they have been shown to be more scalable, efficient and fault tolerant than centralized or other distributed approaches.

P2P classification. Here, we give a brief overview of the P2P classification approaches used for automated tagging in our system. CEMPaR [2] is a P2P classification approach that is based on the cascade Support Vector Machine (SVM) paradigm and uses distributed hash table to reduce the communication cost [2]. First, each peer constructs a non-linear SVM model using its local training data (document vectors and assigned tags). Then, these SVM models (support vectors) are propagated once to one of the super-peers in the P2P network. The super-peers are automatically elected from the P2P network and are located in a deterministic manner, made possible though the use of the DHT-based P2P network. Although the support vectors propagated to the super-peers are in fact the document vectors, they cannot reconstruct the text documents, because only the word ids and frequency information are preserved in the vectors. Hence our system preserves some level of privacy and security. Next, super-peers which collect the local models of peers cascade them to construct regional cascaded models. Documents are then automatically tagged by sending the untagged document vectors to the super-peers, which use the regional cascaded models to predict and assign tags. Tags are then selected and assigned by (weighted) majority voting.

Contrary to CEMPaR, PACE [1] is based on the multiple classifier system, which uses the state-of-the-art linear SVM algorithm to reduce computation and communication cost and tries to improve classification accuracy by adapting to the (test) data distribution. First, peers construct a linear SVM model using their local training data and also perform clustering on the training data. Then, the linear SVM models and the centroids of the clusters are propagated to all other peers. Since no document vectors are propagated to other peers, the system preserves some level of privacy and security. Upon receiving the models and centroids, peers index the models using the centroids (based on locality sensitive hashing). Given a document vector which needs to be tagged, the algorithm retrieves the top k "nearest" models (with respect to the distance between the test data and the models' centroids) from the index. It then predicts the tags using these models, which are in turn weighted according to their accuracy and distance from the test data. Interested readers may kindly refer to the original publications [2, 1] for further technical details.

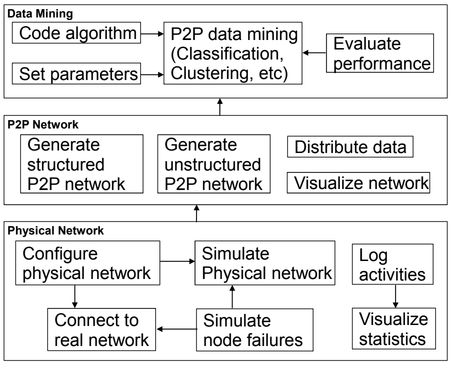
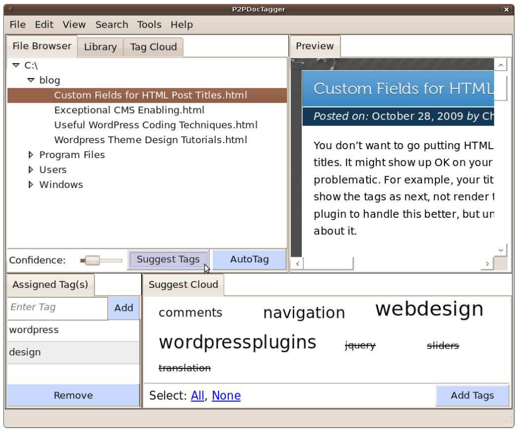


Figure 2: System architecture of P2PDMT.

Tag Refinement. Since tags are assigned by different peers who may have different perceptions of the same document/tag, it would not be unusual to find conflicts in the tag assignments. In such situations, users' intervention would be required. On the discovery of mismatched tags on documents, users can use the tagging interface to modify the assigned tags for the documents. Upon the refinement of tags, P2PDocTagger will automatically update the classification model(s) in the back-end, to adapt to their personal preference for future tagging.

P2P Data Mining Simulation Toolkit. Testing the system under real P2P environments with thousands of peers require a large amount of resources, which we are unable to support. Hence, to test P2PDocTagger under realistic P2P environments, we make use of the P2P Data Mining Toolkit (P2PDMT) that we built on top of Oversim [3]. Oversim is a P2P overlay simulation framework that can realistically model the real world P2P network. P2PDMT extends Oversim providing functionalities to support data mining tasks such as data distribution, algorithm evaluations, result visualization, etc. To the best of our knowledge, the most similar work is the distributed data mining toolkit (DDMT)[3]. However, P2PDMT offers several features not present in DDMT. It is able to simulate realistic P2P environments, including overlay topologies, peer churn, etc. In addition, code written for P2PDMT is reusable in real applications. Figure 2 shows the architecture of P2PDMT. To illustrate how various networking and data mining scenarios can be simulated, P2PDMT allows setting parameters like physical connection of peers, total number of peers in the network, churn model(s), P2P overlay network, training data, size distribution of training data, class distribution of training data, testing data, frequency and timings of evaluations, etc.

[3] `http://www.umbc.edu/ddm/Software/DDMT/index.html`



Custom Fields for HTML Post Titles.html
Exceptional CMS Enabling.html
Useful WordPress Coding Techniques.html
Wordpress Theme Design Tutorials.html

Custom Fields for HTML

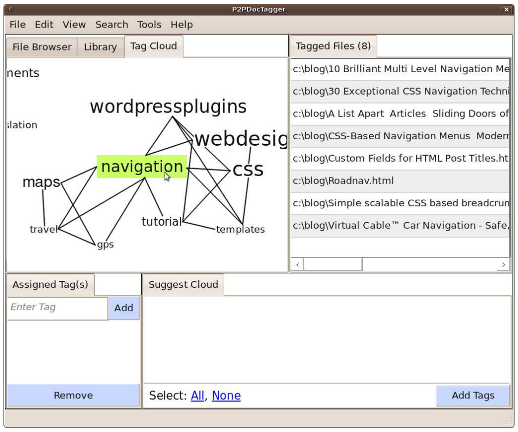
Posted on: October 28, 2009 by Ch

You don’t want to go putting HTML titles. It might show up OK on your problematic. For example, your tit show the tags as next, not render t plugin to handle this better, but un about it.

wordpress
design

comments navigation webdesign wordpressplugins ~~jquery~~ ~~sliders~~ ~~translation~~

Figure 3: Illustration of document tagging.



c:\blog\10 Brilliant Multi Level Navigation Me
c:\blog\30 Exceptional CSS Navigation Techni
c:\blog\A List Apart Articles Sliding Doors of
c:\blog\CSS-Based Navigation Menus Modern
c:\blog\Custom Fields for HTML Post Titles.ht
c:\blog\Roadnav.html
c:\blog\Simple scalable CSS based breadcrun
c:\blog\Virtual Cable™ Car Navigation - Safe,

Figure 4: Illustration of tagged based file browsing.

3. DEMONSTRATION OVERVIEW

We demonstrate P2PDocTagger using real data from `http://delicious.com` collected by Wetzker *et al* [7], which consists of public bookmarks of about 950,000 users retrieved from the site between December 2007 and April 2008. Users with at least 50 (and, to avoid spammers, less than 200) annotated bookmarks were chosen and the corresponding web documents retrieved. 20 percent of the documents with tags are used for training the automated tagger, while tags of the remaining 80 percent documents are removed to be tagged by P2PDocTagger. Our system will be demonstrated on different P2P environments (e.g., DHT-based P2P network with more than 500 peers), and we will show how to setup these different simulation environments for realistic P2P data mining simulations. In the demonstration, we will illustrate how to tag the documents both manually and automatically. Audiences can interact with the system to assign or refine the tags of documents, or browse the documents using the tags.

To illustrate the flexibility of P2PDMT and its ability to simulate real world P2P networks conditions, we will create various P2P scenarios by modifying the network parameters, such as the network size, churn/attrition rate of the P2P network, topology of the P2P network, etc. In addition, we will vary the data distribution on the peers by varying the size and class distributions.

Screenshots of P2PDocTagger are presented in Figure 3 and 4. Users can access the system either through the menu bar (located at the top), or the navigation panel (located at the top left). The three main navigation components are as follows – 1) File Browser, which allows users to browse their file system to tag their documents, 2) Library, where all tagged documents are tracked to allow users to browse or search documents using tags, and 3) Tag Cloud, where files can be accessed through the tag cloud interface. Figure 3 demonstrates how users can add tags to the documents. First, select the documents to tag, e.g., using the File Browser. To automate the tagging, select single or multiple files and press “AutoTag” button. Otherwise, to have the system suggest tags from its distributed and collaboratively trained models, select a single file and press “Suggest Tag” button. Relevant tags will be shown in the “Suggestion Cloud” panel, arranged in alphabetical order, where tags with higher confidence will be in larger font. Low confidence tags can be filtered out (struck out, and placed last) by adjusting the “Confidence” slider in the “File Browser” panel. To add tags to the documents, select the relevant tags and press “Add Tags” button. To manually add tags, under the “Assigned Tag(s)” panel, key in the tag and press “Add” button. Tags assigned to the document are listed in the “Assigned Tag(s)” panel.

In addition, we will demonstrate searching and filtering of documents using the “Library” component and navigation of the documents through the “Tag Cloud” interface. In the Tag Cloud interface presented in Figure 4, tags that cooccur in documents are connected by edges (cf. Figure 3). This provides users with information regarding the tag relationships and captures higher level concepts as presented in Figure 4, where we see two clusters of highly interconnected tags bridged by the word “navigation”. Other uses of the system through the menu bar will also be demonstrated.

4. REFERENCES

[1] H. H. Ang, V. Gopalkrishnan, S. C. H. Hoi, and W. K. Ng. Adaptive ensemble classification in p2p networks. In *DASFAA (1)*, pp. 34–48, 2010.

[2] H. H. Ang, V. Gopalkrishnan, W. K. Ng, and S. C. H. Hoi. Communication-efficient classification in P2P networks. In *ECML/PKDD (1)*, pp. 83–98, 2009.

[3] I. Baumgart, B. Heep, and S. Krause. Oversim: A flexible overlay network simulation framework. In *IEEE Global Internet Symposium*, pp. 79–84, 2007.

[4] E. Cutrell, D. C. Robbins, S. T. Dumais, and R. Sarin. Fast, flexible filtering with phlat. In *CHI*, pp. 261–270, 2006.

[5] W.-T. Hsieh, J. Stu, Y.-L. Chen, and S.-C. Chou. A collaborative desktop tagging system for group knowledge management based on concept space. *Expert Systems with Applications*, 36(5):9513–9523, 2009.

[6] S. Siersdorfer and S. Sizov. Meta methods for model sharing in personal information systems. *ACM Transactions on Information Systems*, 26(4), 2008.

[7] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *ECAI Workshop for Mining Social Data*, pp. 26–30, 2008.