Computer Communications 36 (2013) 191–202

Contents lists available at SciVerse ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Beyond pollution and taste: A tag-based strategy to increase download quality in P2P file sharing systems

Flávio Roberto Santos *, Weverton Luis da Costa Cordeiro, Luciano Paschoal Gaspary, Marinho Pilla Barcellos

Institute of Informatics, Federal University of Rio Grande do Sul, Av. Bento Gonçalves 9500, 91.501-970 Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 2 August 2011
Received in revised form 24 May 2012
Accepted 15 August 2012
Available online 24 August 2012

Keywords:
Peer-to-peer
File sharing
Folksonomy
Tagging
Pollution

ABSTRACT

The degree of autonomy provided to users when publishing contents in P2P file sharing systems allows the dissemination of files with inaccurate, incorrect or imprecise descriptions, either because of the diversity of users' opinions (subjectivity) or due to malice. The lack of proper mechanisms to deal with these issues leads to download of undesired contents, waste of resources (such as bandwidth), and users' dissatisfaction. Current approaches try to identify improper descriptions (e.g., through users' reports) and isolate contents, but such strategies are cumbersome, require considerable time for a "stable view" about the contents to be formed, and therefore promote early and wide dissemination of contents that do not match users' expectations. To overcome this limitation, we propose a novel strategy that allows users to better describe contents and regulates content distribution based on users' perception. The proposed strategy, called DÉGRADÉ, harnesses the power of tags to control the dissemination of poorly, misdescribed contents while it aids the dissemination of sufficiently and correctly characterized ones. Results obtained through a large set of experiments provide evidence about the efficacy and efficiency of the proposed strategy.

1. Introduction

Peer-to-peer (P2P) file sharing systems have been one of the most important Internet applications [1]. The majority of such systems are formed by communities of users sharing different interests and preferences [2]. In each of these communities, users have autonomy to publish, label and describe new contents following their own perception.

As a consequence of such autonomy, it becomes natural the appearance of mislabeled, wrongly described contents. The reason is either the diverse users' perception regarding contents or malice. Concrete evidences and statistics about "pollution[1]" have been widely reported in the recent past (e.g., for BitTorrent communities [3,4]). These evidences are underscored by a previous study [5] and the results of a survey we have carried out with administrators of BitTorrent communities, who informed that around 25% of contents being published are "polluted" and need manual intervention.

To cope with the aforementioned issue, a variety of solutions has been proposed to identify (and marginalize) "polluted" contents. The existing solutions share one important limitation: they are unable to deal with the inherent *subjectivity* to characterize "pollution". Content regarded as "polluted" by a certain user might be considered "non-polluted" by another. Furthermore, Lee et al. [6] have shown that about 70% of users fail to notice certain types of incorrect descriptions. Such fact is neglected in the design of existing approaches, which assume that non-malicious users always detect inappropriate metadata.

As an important step towards increasing download quality in file sharing systems, in this paper we propose DÉGRADÉ, a novel tag-based strategy to control the dissemination of undesired contents. Unlike existing solutions, DÉGRADÉ leverages social tagging system techniques to allow users to express their perception about downloaded contents in a flexible and accurate fashion. Furthermore, and key to our solution, DÉGRADÉ restricts the dissemination of contents when the uncertainty on the vocabulary describing them is high, and promotes their dissemination otherwise. We go beyond the traditional binary polarization ("polluted" *vs.* "non-polluted") of users' feedback in characterizing contents and provide a more effective way for users to express their "taste" – through the assignment of tags that represent their opinion about a content. The control strategy underneath manages to reconcile such "tastes" and downloads. As far as we are aware of, this is the first work to explore the use of collaborative annotations (widely used in platforms such as Delicious, YouTube, and Flickr) within a *systematic* strategy to control content dissemination in file sharing systems.

* Corresponding author. Tel.: +55 51 81719506.
E-mail addresses: flavio.santos@inf.ufrgs.br (F.R. Santos), weverton.cordeiro@inf.ufrgs.br (W.L. da Costa Cordeiro), paschoal@inf.ufrgs.br (L.P. Gaspary), marinho@inf.ufrgs.br (M.P. Barcellos).

[1] In the context of this paper, "pollution" refers to contents published along with inappropriate meta-data.

0140-3664/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comcom.2012.08.012

Our strategy has been evaluated by means of a large set of experiments. We have reproduced BitTorrent swarms to mimic various scenarios of content dissemination, and have employed real traces from Delicious as sequences of tag assignments. In order to measure the "quality" of downloads, we defined a metric based on the variation of the vocabulary describing a content. The experiments carried out showed that DÉGRADÉ was capable of improving this metric. We also reproduced common attacks against file sharing systems, in which malicious users (*i*) disseminate a large amount of undesired contents and (*ii*) hinder users to reach the contents they desire. The results achieved show that DÉGRADÉ is robust even in the presence of malicious users.

It is important to emphasize that the major contribution of our paper is to solve the problem of dissemination of undesired contents among users in file sharing systems, through the instantiation of a tag-based strategy. The value of our contribution is concerned to the challenges that have to be addressed when carrying out such an instantiation. Our choice for social tagging systems, in turn, is due to their inherent property of enabling users to freely express their subjective opinion (through annotations) about a given content. As shown in a previous study [7], these systems already handle the problem of subjectiveness.

The remainder of this paper is organized as follows. Section 2 discusses some of the main strategies to tackle content pollution, and describes the use of tags in the context of tagging-based systems. The proposed strategy and equations that govern its behavior are presented in Section 3, whereas evaluation results are discussed in Section 4. Practical aspects related to the instantiation of our mechanism in P2P systems are discussed in Section 5. Finally, Section 6 closes the paper with concluding remarks and prospective directions for future work.

2. Related work

The strategy presented in this paper combines two distinct fields: pollution control in file sharing systems and tag-based systems. In this section we discuss publications related to these areas.

2.1. Pollution dissemination control in file sharing systems

There has been substantial investigation on solutions for controlling the dissemination of polluted contents within the peer-to-peer research community (please refer to Hoffman et al. [8] for a research overview). In general, these may be categorized in one of the following classes: feedback-based and moderation-based solutions.

Credence [9], Scrubber [10] and Hybrid [11] represent some of the most representative examples of user feedback-based reputation systems. Credence relies on the assumption that honest peers issue similar votes when establishing the authenticity of a given content. In this case, peers having higher voting similarity are more trustworthy than others having lower similarity. Scrubber, similarly to Credence, adopts a more aggressive approach to penalize polluters. Hybrid, in turn, combines peer and content reputations, thus managing to penalize peers that only occasionally upload polluted content. An issue common to these solutions lies in their vulnerability to *whitewashing* attacks [12], since newcomers and newly published contents start with a neutral reputation. To tackle this issue, in a recent work our research group has proposed a generic *conservative* pollution control strategy [13] and instantiated it as a mechanism for BitTorrent communities [14]. The strategy is conservative in which it throttles the dissemination of newly published content, and allows the dissemination rate to increase according to the proportion of positive feedback issued about the content. It avoids the problem of pollution dissemination at the initial stages of a swarm, when insufficient feedback is available to form a reputation about the content.

The second class of solutions, which are moderation-based, relies on human intervention for accepting or rejecting content submissions. Solutions vary according to the degree of decentralization. On one extreme, administrators inspect every published content individually. On the other, users may issue comments or even denounce some content being shared in the community. Centralized moderation-based approaches need to rely on a limited number of system administrators. This imposes a severe constraint to the rate in which content is evaluated, a problem for communities with higher content publication rates.

In summary, despite the clear advances towards addressing the issue of content pollution in file sharing systems, existing solutions do not take into account the inherent *subjectivity* behind the evaluation of shared contents. In the next subsection, we describe how tags are employed to tackle this problem in the context of collaborative systems.

2.2. Tagging-based collaborative systems

User-specified tags have emerged as an alternative, or complementary, way of describing online resources such as web pages, photos, and videos. Since its inception, tags have been largely explored in collaborative platforms such as Delicious, YouTube, and Flickr. In line with their growing popularity, the also called social tagging systems have received great attention from the scientific community. Examples of research directions under investigation include understanding the semantics of tags associated to different types of resources [15], and the study of their structure and dynamics [16].

It is consensus in the literature that tags substantially improve searching and retrieving of resources on collaborative platforms, and have the potential to boost reputation systems [16]. Folksonomies add an additional layer of descriptive information about the content, and enable dealing with the subjectivity associated with the process of resource description. In this realm, Andrade et al. have proposed a peer-to-peer annotation approach, where users can freely annotate available resources [17]. Despite its potential, one of the main problems is the malicious or accidental assignment of inaccurate/incorrect tags to resources. This behavior, known in the literature as "tag spamming", has been widely discussed in recent research [18–20].

In this paper we leverage the power of tags to control the dissemination of poorly, misdescribed contents and aid the dissemination of sufficiently and correctly characterized ones. There are two challenges related to the adoption of tags as a mechanism to deal with subjectivity in P2P file sharing systems: identifying when the *tag cloud* effectively reflects the nature of a given content and, as just mentioned, dealing with tag spamming attacks. The following sections detail how these challenges are addressed by the proposed strategy.

3. DÉGRADÉ model

This section introduces the downloads control strategy supported by content vocabulary variation. This strategy, called DÉGRADÉ, can be divided in two main components. The first one is the *tagging component* used by users to type in the set of tags associated to some content. The second one is the *aggregator* service, where the tags are stored and employed to build our solution. The tagging component allows users to annotate contents similarly to the popular Delicious bookmarking system. After each tag assignment, the aggregator service updates the content vocabulary and the variation metric used to control further requests for