Tag-Based Navigation for Peer-to-Peer Wikipedia *

Jenneke Fokker
Dept. of Industrial Design
Delft University of Technology
The Netherlands
j.e.fokker@tudelft.nl

Johan Pouwelse
Dept. of Computer Science
Delft University of Technology
The Netherlands
j.a.pouwelse@tudelft.nl

Wray Buntine
Dept. of Computer Science
Helsinki Institute of
Information Technology
University of Helsinki, Finland
buntine@hiit.fi

ABSTRACT

We introduce *P2P Wikipedia*, a prototype of a personalized tag-based navigation system for Wikipedia multimedia content. It is the first peer-to-peer (P2P) file sharing system able to deal with large files like movies, music, and software, but that is also scalable to HTML content. The combined techniques in our prototype are the automated calculation of tags from HTML content, a personalized P2P file sharing system built on a social network, the use of incentives for user cooperation to optimize system performance, and the design of a user interface with advanced navigational features.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

General Terms

information retrieval, metadata management, peer-to-peer

Keywords

tagging, peer-to-peer, Wikipedia, incentives, wisdom of crowds

1. INTRODUCTION

Wikipedia is a web-based encyclopedia, written and edited collaboratively by volunteers (`wikipedia.org`). Its popularity has led to excessive and costly bandwidth usage. This makes it undesirable - if not impossible - to augment pages with extensive video footage. The multimedia version is not able to scale to Gigabyte files either (`commons.wikipedia.org`). P2P technology presents the solution for distribution of Wikipedia content. It can reduce the hosting costs and enables the integration of large multimedia files. There are many methods to search efficiently in text based files with keywords. But this is not that straightforward for video files. Apart from known metadata such as director, title, genre, actors, and year, it is hard to extract keywords from video footage automatically. That is why voluntary tagging is ideal. To illustrate this, consider finding a particular movie, but you have forgotten the title, the names of the actors, or any other metadata that could have helped finding the movie directly. All you remember is that it involved a Citroën DS and a Japanese man. Keyword searching would not lead you directly to *The Goddess of 1967*. But when many users have tagged this movie freely and massively, the chance is much bigger that some have used the tags *Citroën DS* and *Japanese man*, and consequently the chance is also bigger that you will find the movie.

We believe that tags are an augmentation to keyword searching in video files. They facilitate associative searching and increase the possibility of serendipitous content discovery from the Long Tail [2]. We define a tag as a freely chosen descriptor, or label which refers to one aspect of an object. The tag cloud has recently emerged as a popular navigation method through large amounts of tags. The cloud is a representation of the frequency-based relation of tags. An example of a CiteULike tag cloud is shown in Figure 1.

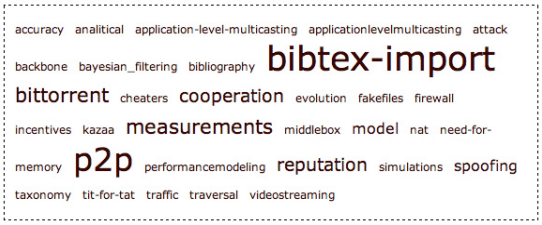

Figure 1: Tag cloud (source: Wikipedia.org)

Essential for good system performance is that users are willing to tag voluntarily. Especially uncommon tags describing specific features in movie scene, e.g. the *Citroën DS* facilitate richer navigation. The two broad questions we address in this paper are how to organize and store information in a scalable and efficient manner, and how to stimulate more users to tag more. As a case study we augment Wikipedia with tags and P2P technology. We believe our approach is generic and can be applied to information organization and storage in general.

In this paper we present three contributions. First, a generic operational tool to automatically calculate tags from Wikipedia articles: `Text2Tag`. It bootstraps the generation of tags and stimulates users to participate in tagging. Second, a fully functional P2P file sharing system with a scalable architecture for personalized sharing and real time streaming of large files, like video content: `Tribler` [10]. We are expanding Tribler with tag-based semantic clustering. Third, a user interface with incentives to cooperate, and efficient navigation through large amounts of tags.

*(Produces the WWW2006-specific release, location and copyright information). For use with www2006-submission.cls V1.4. Supported by ACM.

Copyright is held by the author/owner(s).
WWW2006, May 22–26, 2006, Edinburgh, UK.

2. TAGS

Searching and browsing through near endless amounts of texts, web pages, or accross the web is a serious challenge. A problem is the lack of high-quality metadata to enable efficient search. This has been described accurately in [4], stating that the past decade has shown that metadata are often unavailable and searching is best done using the text itself rather than relying on metadata. To create good quality metadata, one would need professionals to do so. Free and voluntary tagging is a good alternative. But, prerequisite for success is large scale tagging. Websites such as `Flickr.com`, `CiteUlike.org`, and `Del.icio.us` have shown the popularity of tags for search and attracted millions of users. The key is their use of volunteers to augment content with tags. Every visitor to such websites can participate in this collaborative categorization. Different websites may have different underlying reasons why people tag (e.g. see [7]), but the altruistic effect is helping other users in exploring the content. The system performs better if more users participate.

However, tags can suffer from ambiguity and arbitrariness. This can be illustrated by the indexing process in a library: A professional librarian has strict rules to abide by and all library members can rely on the integrity of the indexing. But when everyone who wishes to can contribute to indexing books and there are no immediate rules, it's obvious that ambiguity and arbitrariness will arise. Everyone will index in a way that makes sense to themselves. The ambiguity of tagging - as described in [6] - is illustrated by statistics from `Flickr.com`. Table 1 shows the result of our measurement of the various synonyms for the US city of New York.

Table 1: Synomyms of a Flickr tag (Dec'05)

Tag	Number of Photos
nyc	340,000
newyork	228,000
newyorkcity	106,000
new-york	61,000
new-york-city	13,000
ny	67,000
bigapple	2,000

Even though there seem to be initial problems of scalability, and ambiguity of tagging systems, 'with sufficient critical mass, truth would arise from consensus' [16], also known as the *Power of Collective Intelligence*, or the *Wisdom of Crowds* [13]. The advantage is that it can facilitate the task of finding popular tags, and stimulate serendipitous exploration of the tagged universe. The quality of tags in *P2P Wikipedia* is a result of their quantity and the fact that we let people moderate each other in a wiki-style. This is a proven concept: Wikipedia is said to be in the same, or sometimes even higher, league as the Encyclopedia Brittanica [5]. We believe the biggest challenge is stimulating users to tag content. Our approach to this is exploiting social phenomena, as will be explained in Section 4.1. Furthermore, we ensure a bootstrap for tagging by implementing smart algorithms in our software discussed in the next section.

3. TAGGING THE WIKIPEDIA

The Wikipedia collaborative encyclopedia is our tagging case-study. It is chosen for its availability of content and embedded links, and its large user-base. This section describes relevant aspects of Wikipedia and how we generate tags from the Wikipedia database dump using our `Text2Tag` toolset. We are well-aware of the difference between tags (which are by definition user-generated) and calculated tags (which should in fact be called keywords), but we choose to call them the same because they can be mixed when the use of *P2P Wikipedia* progresses. For calculation, the tags we use, nevertheless, have been inserted by authors, but as links, not tags, thus they are not entirely computer generated. Throughout this paper we will clarify the nature of tags: either user-generated or calculated.

Versions of Wikipedia are available in many different languages. The English language version is the largest with over 930,000 articles in December 2005 with approximately 4.5Gb of uncompressed text (HTML removed) and 580,000 image files including 28.000 with scalable vector graphics. One can perform a targetted search using a Lucene[1] system. Wikipedia also offers topical information on current news daily as well as portals such as the Science Portal. Wikipedia consists of pages with a unique topic name, which can be seen as a unique tag. For example, pages exist for 'democracy', 'coal_mining', and 'Cultural_elements_of_Buddhism'. However, a single Wikipedia page can describe numerous subtopics and describe numerous facets of the main topic, and thus cover multiple tags.

Table 2: Ambiguity in Wikipedia pages

Tag	Wikipedia page topic
analytical_engine	Bruce_Sterling
analytical_engine	steampunk
analytical_engine	alternate_history
analytical_engine	The_Difference_Engine
ananda	Cultural_elements_of_Buddhism
ananda	History_of_Buddhism
ananda	List_of_Buddhist_topics

Table 2 shows some examples of the relation between tags and Wikipedia page topics. In December 2005 Wikipedia included roughly 32,000,000 links between pages, and 790,000 redirects from variant topic names (e.g., 'Abel' to 'Cain_and_Abel'). Moreover, the link text associated with tags, the text an author has written for a link, is every bit as varied as the tags in Flickr or other systems. Thus our association of the link to a tag means we have, in effect, had the synonym problem solved for us.

We developed software to generate tags from Wikipedia as a bootstrap for user-generated tags. These tags are the title text for pages that are not disambiuation pages or stubs, that have more than 4 in-links (where 4 is arbitrarily chosen), and that may also be category pages. A page containing a link (possibly through a redirection) to such a page is said to contain the 'tag'.

The challenge is not only generating tags, but also organizing them into top-tags, sub-tags, subsub-tags, and adding weights. We implemented the generic `GenerateTopTags` function to generate tags. This function can generate both top-tags, sub-tags, subsub-tags, and handle the `AND` operator. It increases freedom searching and exploring content, and this bootstrap should stimulate more users to generate more tags for their own content.

[1] `http://lucene.apache.org`

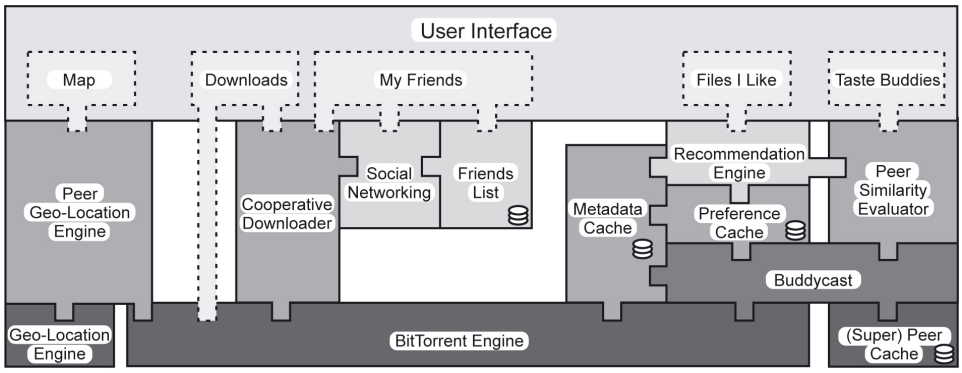

Figure 2: The system architecture of Tribler

Candidate subtags and matching documents are generated using both an inverted and a forward index (i.e., bags of tags for each document). Top-tags are ranked using a PageRank score [8] that downgrades the time tags (time is richly tagged in Wikipedia, so tends to dominate other concepts). Ranking a candidate list of sub-tags is done by combining a standard `idf` score [3] relative to the tag. `Idf` is inverse document frequency, and the sub-tags frequency is taken from the set of documents having the major tag, not the full collection. Ranking a candidate list of documents matching a particular query made up of tags is done again with a standard `tf*idf` score.

4. NEXT-GENERATION P2P

We are currently working on generalizing P2P file sharing to supporting content distribution in general. For this we implemented an operational P2P system called *Tribler* [10]. It is based on the popular Bittorrent protocol. Source code is available from `Tribler.org`. We have merged real-time MPEG4 streaming with Bittorrent [11]. Furthermore, we are merging our file sharing system with web technology, thus creating a decentralized Wikipedia. Several possible business models exist, the donation-based approach exemplified by Wikipedia donation rallies, a model where targetted advertisements pay for the hosting costs, and variations to these themes, e.g. `Amazon.com`. We show another model where users index the content, moderate existing content, and provide the resources for persistent storage, publication, and distribution.

4.1 Tribler

In this section we present the architecture of our Tribler social-based P2P file-sharing system, which is built on top of the Bittorrent protocol. Figure 2 depicts the architecture of the Tribler network client. Rectangles represent client modules. The extrusions represent *make-use-of* relationships. To achieve backward compatibility with the existing Bittorrent network, while offering our users extended functionality, we only made modifications and extensions to the Bittorrent client software. Our system is based on the ABC open-source client [1]. By extending this popular client we aim to have a large user base in a relatively short time, besides having a tested code base for our implementation.

Social phenomena The prime social phenomenon we exploit in Tribler is an analogy to evolutionary biology: kinship fosters cooperation [9]. This kinship is interpreted as friendship or belonging to a community, because genetical relations are not taken into account. Similar taste for content can be one of the foundations for an online community with cooperative behavior, instead of remaining an ad-hoc group of non-cooperating strangers. In order to create effective social groups in Tribler, we use an approach that stimulates the ability to distinguish friend and newcomer from foe. For this, we de-anonymize peers by having every user choose a permanent identity, and facilitate the actual creation of social groups. Tribler transfers user nicknames between users automatically. The *Social Networking* module in Figure 2 is responsible for storing and providing information regarding social groups (the group members, their recently used IP numbers, etc.). This will be discussed in the next paragraph.

Tag-buddy based content discovery Locating content is critical for P2P systems. Current solutions are based on one or a combination of query flooding, distributed hash-tables, and semantic clustering. We take a next step by connecting *people* with similar tastes instead of focusing on *files*, and by using full metadata replication. In Tribler we exploit the fact that people with similar tagging behaviour - also known as *tag buddies* - have related taste.

Using the *Files I Like* module, each peer indicates its preference for certain files and their associated tags. By default, the preference list of a peer is filled with its most recent downloads. We have developed an algorithm called *Buddycast* that employs an epidemic protocol [15] to exchange preference lists using the overlay swarm and that can efficiently discover a user's tag buddies. The *Peer Similarity Evaluator* module in Figure 2 is able to compare similar preference lists.

4.2 P2P Wikipedia

Our software will enable the extension of Wikipedia with multimedia and tags. We are using our `Text2Tag` toolset to import Wikipedia tags into Tribler. Due to the excessive Wikipedia bandwidth usage it has not been possible to augment pages with extensive video footage. The required servers and Internet connection can not be supported by the Wikipedia donation-based approach only. But by integrating the proven Bittorrent technology we can reduce bandwidth bottlenecks and create a scalable system.

There are four key extensions needed on Tribler for *P2P Wikipedia*. First, the ability to display Wikipedia content inside Tribler, thus add an embedded web browser. Second, remove the Bittorrent tracker from the content discovery

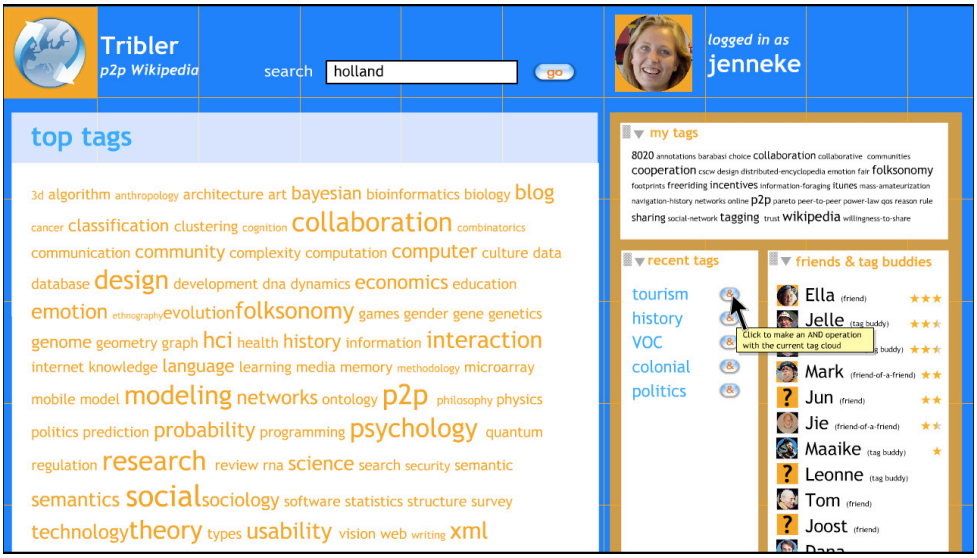

Figure 3: Main navigation screen

architecture using epidemic protocols [15]. Third, eliminate the need for .torrent files by using Merkle hashes and embedding these into a URL. Fourth, build a version management system on top of the Bittorrent content layer to enable collaborative editing.

5. INDUCING COOPERATION

In this section we explore incentives to induce cooperation. In a P2P system it is important for users to voluntarily cooperate. In *P2P Wikipedia* maximal cooperation is essential, because the success of tagging depends on their quantity. Voluntary cooperation does occur in the online world, but we believe we can stimulate more users to cooperate. By using the right incentives we hope to induce users to cooperate and explore the 'tagged universe'. Furthermore, the user interface has to be an attractive representation of and supporting the ideas about inducing human cooperation.

Incentives to cooperate Social psychologists identify two basic motivational forces [14] for cooperative behavior. First, *instrumental, or environmentally driven motivation*: people either see the chance of a reward if they cooperate, or fear punishment if they do not cooperate. For instance, websites like (`amazon.com`) reward their users with *badges* that show their status and higher ranks in *top-n lists*. Second, *internally driven motivation*: the influence of personal values in the form of obligation to the group and its rules (legitimacy), and of attitudes relevant to the group (commitment, satisfaction, feelings toward group authorities, and loyalty). The latter is also known as ingroup identification [12]. How people behave within their group (or virtual community) is also influenced by their wish to create a positive public self. Factors that influence the way people present themselves positively include the willingness to cooperate, the feeling of belonging, competitiveness, the need to distinct oneself from others, the possibility to convince others of ones opinion or taste.

Personalized navigation Tag-based navigation can be an additional way to explore Wikipedia, especially for video files. A user's taste is learned from creating new content, tagging, moderating, searching, and browsing. The *Buddycast* algorithm then calculates the user's taste buddies - also known as *tag buddies* in *P2P Wikipedia* - to create a sense of belonging to a community. From all this it will also be easier to calculate recommendations, as is done in the *Tribler* system, and tag-to-tag similarity. They both result in a much richer and more serendipitous exploration of the 'tagged universe'.

The main navigation screen in Figure 3 shows the calculated tag cloud with 50 - 100 top Wikipedia tags on the left, arranged alphabetically. Their relative ranking is expressed in font-size. The means of navigation in this prototype are keyword searching with one or two keywords, and tag-browsing. The area on the right is used for personalized settings, containing:

a) My Tags. Summarizing the user's most often used tags. Like the tag cloud on the left, this personal tag cloud is arranged alphabetically, and the ranking is expressed in font-size.

b) Recent Tags. A list of most recently viewed tags (xn-1...n-12). The user can perform an AND operation of one of these tags with the currently viewed tag cloud by clicking the button behind that tag.

c) Friends and Tag Buddies. An overview of friends, friends-of-a-friend and buddies that are currently online. Information about a user's friend or friend-of-a-friend comes from the integration of an existing social network (not yet built in the prototype). Showing this social network will de-anonymize the system and stimulate contribution. Note that users are rewarded with stars for their cooperation.

Figure 4 shows the tag cloud resulting from the search operation 'holland AND tourism' and the directly matching Wikipedia results. Floating the mouse over a tag brings about two things (see Figure 4). First of all, directly related tags in the tag cloud are highlighted. The tag *amsterdam* has a number of directly co-occuring tags, e.g. *schiphol* and *rain*. And secondly, information about the tag is shown in a tooltip. The tag *amsterdam* in this figure represents the category *amsterdam* with a number of sub-tags and Wikipedia articles. The colored bar in the tooltip shows how much the tag relates to the two keywords relatively. Both the high-

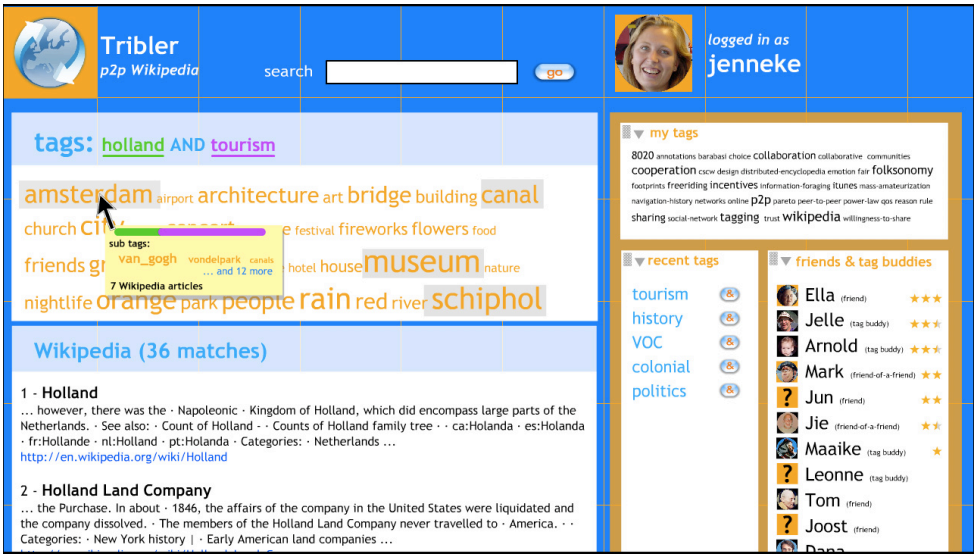

Figure 4: Mouse-over showing sublevel

lighting and the tooltip facilitate more efficient navigation, because information is available in advance.

6. CONCLUSIONS

We have created a way of efficient searching within video content. In this paper we explained why, when, and how tag-based navigation can augment traditional keyword searching for `Wikipedia.org`, and why cooperation is needed from the users to do so. We have also described how excessive and costly bandwidth problems can be solved by using P2P technology, which accumulates distributed resources and can also be applied to web content. We presented our vision for the design of the user interface which incorporates incentives to cooperate.

Currently, Tribler is available for download from `Tribler.org`. We are planning controlled experiments in our laboratory with the user interface on a few dozen users in the second half of 2006. This will enable us to test and improve the efficiency of the incentives to cooperate and tag-based navigation. We are currently also working on a taxonomy of cooperation inducing interface features, that will provide guidelines for the user interface design of systems depending on user participation. Furthermore, our ambition is to merge and integrate all information from `Wikipedia.org`, `Del.icio.us`, and `CiteULike.org` into a single coherent P2P system with tags as the organizing principle.

7. ACKNOWLEDGEMENTS

The authors would like to thank Piet Westendorp and Huib de Ridder from the Delft University of Technology, and the Tribler developers for their contributions to this paper and the software.

8. REFERENCES

[1] `http://sf.net/projects/pingpong-abc`.

[2] C. Anderson. The long tail. *Wired Magazine*, October 2004.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[4] D. C. A. Bulterman. Is it time for a moratorium on metadata? *IEEE MultiMedia*, 11(4):10–17, 2004.

[5] J. Giles. Internet encyclopaedias go head to head. *Nature*, December 2005.

[6] M. Guy and E. Tonkin. Folksonomies: Tidying up tags? *D-Lib Magazine*, 12(1), January 2006.

[7] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4), April 2005.

[8] A. Langville and C. Meyer. Deeper inside PageRank. *Internet Mathematics*, 1(3):335–400, 2004.

[9] E. Pennisi. How did cooperative behavior evolve? *Science*, 309(5731):93, July 2005.

[10] J. Pouwelse, P. Garbacki, J. Wang, A. Bakker, . J. Yang, A. Iosup, D. Epema, M. Reinders, . M. van Steen, and H. Sips. Tribler: A social-based peer-to-peer system. In *5th Int'l Workshop on Peer-to-Peer Systems (IPTPS)*, Feb. 2006.

[11] J. Pouwelse, J. Taal, R. Lagendijk, D. Epema, and H. Sips. Real-time video delivery using peer-to-peer bartering networks and multiple description coding. In *IEEE Conference on Systems, Man & Cybernetics*, October 2004.

[12] S. A. Reid and M. A. Hogg. Uncertainty reduction, self-enhancement, and ingroup identification. *Personality and Social Psychology Bulletin*, 31(6):804–817, June 2005.

[13] J. Surowiecki. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* Anchor Books, 2005.

[14] T. R. Tyler and S. L. Blader. *The Antecedents of Cooperative Group Behavior*. Psychology Press, 2000.

[15] J. Wang, J. Pouwelse, J. Fokker, and M. Reinders. Personalization of a peer-to-peer television system. In *4th European Interactive TV Conference (EuroITV)*, May 2006.

[16] A. Weiss. The power of collective intelligence. *netWorker*, 9(3):16–23, 2005.