Content organization and discovery: state-of-the-art and new ideas for P2P-Fusion

Zoltán Prekopcsák
pz559@hszk.bme.hu

May 28, 2007

Essay for the *P2P-Fusion* research project at
Budapest University of Technology and Economics

Abstract

This paper is a state-of-the-art report of content organization and discovery for the P2P-Fusion research project. It has an extensive list of tools and examples, mainly focusing on social tagging and recommender systems. Beyond the examples, it discusses new trends and theories in this area and gives a basic specification for a social p2p application.

Keywords: content discovery, content organization, tagging, tagclouds, recommender systems, collaborative filtering, P2P-Fusion

Contents

1 Introduction

Researchers and companies have a long-time interest in efficient methods of information filtering. 50 years ago the later Nobel Prize-winner Herbert Simon created the term bounded rationality to describe how people make decisions in areas with too much information - and that was many years before the internet (Babarczy et al., 2004).

As we have seen the boost of available information with the spreading of the internet, we have seen the same effect with audiovisual content in the past few years. Nowadays, all new mobile phones has built-in support for making home videos, which means that nearly everyone can create audiovisual content and share it on sites like YouTube or Google Video. Statistics show that in August 2006 there were more than 6 million videos on YouTube (Gomes, 2006), which includes a significant amount of user-generated content. It is impossible to handle this amount of content without intelligent methods for organizing and discovering items.

Search technology made it possible to search and navigate on millions of web pages, but audiovisual content usually doesn't have textual description attached, and it seems that creating automatic transcript of an audio track is a hard problem to solve. Expensive business applications offer solutions for certain languages and domains, but they won't be able to fully capture the essence of audiovisual files in the next few decades. We need efficient tools to collect and manage metadata for audio and video files, so we don't get lost in the catalog of items. A group of web companies have innovative ideas on how to solve this problem and they are often called Web2.0 or social web. They make it easier to search, browse and navigate between items and often they even make automatic recommendations. In the first part of this paper I will analyze these websites to collect the most successful ideas and tools.

Another problem with audiovisual content is that it's quite expensive to host and transfer these files. Though the prices are falling rapidly, the number of videos are rising faster, and we should also mention the improvement in quality which obviously causes growing file sizes and growing costs. Peer-to-peer (p2p) applications offer a great alternative by sharing space and bandwidth between the users. These p2p systems are more complex and harder to design and implement than websites, so they lack a lot of functionality that is already available on the web. Lately, there are a few initiatives to implement these tools into p2p systems. Joost brands itself as the future

of television and Azureus Vuze is one of its competitors, both supporting professional television channels and user-generated content sharing too. A third initiative, called Tribler, is the joint effort of two Dutch universities. Tribler's aim is to create a social-based p2p file sharing application (Pouwelse et al., 2006).

P2P-Fusion is making another step further by supporting creative reuse of audiovisual content and many social enrichment mechanisms. Fusion is focusing on communities creating and sharing audiovisual content.

The scope of this document is to look through content organization and discovery techniques used in social web applications and proposed in the most recent research papers. In the second part, I outline a possible specification of content organization in P2P-Fusion and describe the differences with the current state of other p2p video delivery initiatives.

2 State-of-the-art

There are many different content organization methods known from library and archival sciences like catalogs directory trees and classification systems. Although these methods have been working well for many years, they lack an important property: scalability. Content items on the internet are showing an exponential growth, which means the same tendency for the need of manpower and money when using these techniques. Centralized content organization is unable to keep up with this growth.

As content items tend to be user-generated, growth in the amount of content means a similar growth in the number of users. Hence, these users can keep up with the growth, so they can organize the content in a collaborative way. Although it seems to be the perfect solution, it has some serious drawbacks. Collaborative methods are always vulnerable to spam and vandalism, which makes it less reliable. Furthermore, it's hard to find an incentive for participation. Most of the content organization work is done for personal use rather than public benefit (Golder & Huberman, 2006), but the work done for personal use might be useful for others too.

Another drawback is that collaborative methods usually need to reach a critical amount of users before they create a real public benefit. There are many different methods, but some of them are only present in research papers, because they haven't reached the critical mass yet. In this section, I will describe successful content organization methods used in real applications, and the next section will be about new theories and trends that might become even more useful.

2.1 Tagging and folksonomies

Tagging is a form of classification where users assign tags to entities. Unlike categories, tagging is unconstrained, users can tag an entity with whatever they feel relevant. Tagging is typically used in dynamically changing areas, like the internet.

This system of organization was called "folksonomy" by Thomas Vander Wal by combining the words "folk" and "taxonomy" (Smith, 2004). While usual taxonomies are built top-down, folksonomies have no hierarchy or internal structure. Data mining techniques can be used to discover related tags and assumed hierarchy, but tagging has no explicit structure. Of course, it has advantages and drawbacks too.

2.1.1 Advantages and drawbacks

The main advantage is best shown by a survey report, which states that 28% of Americans have added tags or categories, and 7% add tags daily (Pew, 2007). While content organization was done by specialists in the past, now tagging makes it possible for everyone to participate and share the work. Assigning a tag is as easy as typing a word and pushing a button, so there is no need to study library science before. The harder question is: what is the incentive to participate in collaborative content organization? The first is that tagging keeps our collection organized. We usually don't care about the public benefit, we just want to maximize ours, so we tag important items to be able to find it later. Second, we usually tag items that we like or find important, so we create tags to promote it and make it easily reachable for others. Some researchers argue that there are other forms of motivations like expression, performance and activism, but these are less common than the above two (Zollers, 2007).

This model sounds great, but it has some defects that we should be aware of. Of course, as any collaborative method, it's vulnerable to spam and vandalism, but there are issues that are present with normally behaving users too. Homonyms (different meanings for the same word) cause failure in precision, which means that some of the results will be completely irrelevant. Synonyms (different words for the same meaning) and word inflection (like plural forms) harm recall, so we won't find all of the relevant items. Although these phenomena significantly harm user experience, tagging is still one of the best choices to organize dynamically growing collections of content.

2.1.2 Folksonomies

Although tagging is a relatively simple idea, there are many different realizations of it. The two main categories are broad and narrow folksonomies. It's better to describe them by two popular examples. Delicious (http://del.icio.us) is a social bookmarking service where you can store links to your favorite websites and tag them. These tags are shared with the other users, so others can find an item by the tag that you assigned to it. Everyone can add tags to any items, so the same tag can be added by more people, which produces a power law curve as the distribution of tags (Vander Wal, 2006). This is called broad folksonomy.

On the other hand, Flickr (http://www.flickr.com) has a different ap-

Site	Folksonomy	Tagclouds	Aggregator
Youtube	narrow	no	no
Flickr	narrow	yes	no
Delicious	broad	yes	yes
Last.fm	broad	yes	no
Technorati	no	yes	yes
Blogtelevision	broad	no	yes

Table 1: Tagging tools of social websites

proach. Pictures on the site can only be tagged by the uploader and those that he/she allows to. Tags are singular, one tag can be assigned to an item only once. Compared to Delicious, tagging is done for popularizing the work, not for personal organization. These tags tend to be more descriptive, while broad folksonomies have tags like "wishlist", "toread", "cute" and other types of subjective keywords (Mathes, 2004).

2.1.3 Other examples

There are different tagging solutions that don't fit into these two categories. The most basic form of tagging can be seen in blogs, where only the creator adds tags to the content. Technorati (http://www.technorati.com) and other services aggregate these tags, so users can search and browse them. As an item is tagged by only one person, it doesn't produce a folksonomy, but enhances search in blogs. Slashdot (http://www.slashdot.org) is experimenting with moderated tagging, which seems to be an interesting blend between broad and narrow folksonomies. Blogtelevision (http://www.blogtelevision.net) is an online video aggregator, which lets users tag and comment any video available on other video sharing sites. There are many other services that allow tagging, but they are more or less similar to the ones mentioned above.

2.2 Recommender Systems

While tagging is useful for browsing, searching, and navigating, there is an automatic method that recommends new content items for the user. It can propose an item that we would never think to create a search query for (for example, because we are not familiar with the correct keywords). It also helps to discover relevant items when we don't have the time to browse through the catalog.

Recommender systems analyze user profiles, content items, and the connections between them, and try to predict future user behavior. The idea is "based on the heuristic that people who agreed in the past will probably agree again" (Resnick et al., 1994). The result usually appears as a list of recommended items what the user may like. It is used for marketing in e-commerce sites, finding relevant content in audiovisual archives, and many more. Recommender systems are widely considered as an important part of the emerging social web (Riedl, 2006).

2.2.1 History

It is hard to determine the exact date of the first appearence of a new idea, but we can say that the first recommender systems were born in the beginning of 1990s. The Tapestry document filtering system is considered to be the first to use collaborative filtering method in 1992 (Goldberg et al., 1992). In this system, users could create different filters for incoming mails and netnews. In 1994, the GroupLens research project of University of Minnesota created an automatic recommender system for UseNet news (Resnick et al., 1994). The algorithm could recommend news based on the ratings of others.

At the same time, Upendra Shardanand was writing his thesis at MIT about using recommender systems for music. Along with his supervisor, Pattie Maes, they created a personalized music recommender system, called Ringo. It was working on an e-mail interface and reached a userbase of 2000 in a very short time (Shardanand & Maes, 1995).

Grouplens developers created the basics of the first famous recommender system, which started operating in 1997 for the Amazon.com online bookstore. The recommended items appeared on every product page under the famous expression: "Customers who bought this item also bought". These recommendations were automatic and easily understood by the users, but nowadays Amazon also provides personal recommendations (Linden et al.,

2003). They have an ongoing active research on the topic.

In the last few years, recommender systems became an essential part of online shopping, and we see more and more startups for music, movie and news recommendations. It is an interesting task for the future to explore more possible areas of application.

2.2.2 Input data

Recommender algorithms needs huge amount of data that describe user profiles, content items and connections. Data can be collected in different ways. Explicit (or active) data collection needs the active participation of the user. It means that the user has to evaluate some items before the system starts to generate recommendations. In most cases, the user is asked to rate or rank items. For explicit data collection, incentives are really important. Every action should have an effect on the system, which is visible for the user. It can change the user's profile, give new recommendations, or present other effects, so the user feels that there is a useful result of his/her action. If the recommendations are not real-time and the user sees no effect, then there will be no incentive to keep on rating.

Implicit (or passive) data collection is completely hidden from the user. It relies on usual actions, like purchasing or viewing the item, listening to a music and so on. These actions have old and stable incentives and teaching the recommender is just a side-effect. The main advantage of implicit data collection is that it can collect much more data than with the explicit way as it doesn't need any specific action from the user. On the other hand, it has some problems too. Implicitly collected data is sometimes unreliable, because the item could be viewed by a friend or it was purchased for someone else, so it doesn't represent the taste of the user. To solve this problem, implicit methods are sometimes using secondary explicit input, which is very similar to information theory's relevance feedback (Salton & Buckley, 1990). It means that users can give explicit feedback for the system about the relevancy of the recommendations, so it has a backwards effect on the importance of collected data items. Certain systems let users manually edit the collected data, but this method needs great understanding and high activity from the user.

2.2.3 Algorithms

There are many different algorithms for recommending content items, but the most important approach is collaborative filtering. It uses rating (or other relevance metric) as input data and tries to predict future ratings of unrated items. The main advantage of CF is that the same algorithm can be used for books, music, movies, as it only uses the data about connections between users and items. The two classes of CF algorithms mark the researchers of two different disciplines working on the same problem. Specialists of information retrieval and statistics tend to use memory-based algorithms, while people from the machine learning discipline are mostly using model-based methods.

Memory-based algorithms use all previous ratings (or other data) for the prediction of the new one. Classic solutions define user similarity metrics and predict new ratings based on the ratings of most similar users. There are many different solutions and enhancements for the prediction formula, but the efficiency of the algorithms depend on the dataset, so there is no clear answer for which is the best one. Lately, many researchers proposed that algorithms should define item similarities instead of the users. They state that these algorithms are more scalable and mostly create higher-quality recommendations (Sarwar et al., 2001). Amazon is also using item-to-item collaborative filtering with 29 million users and several million products, so it is an impressive proof of concept (Linden et al., 2003).

Model-based methods have a slightly different approach. The previous ratings are used to teach the model and the predictions are made afterwards. Different solutions use different machine learning techniques for the model, but all try to make simplifications for rating predictions. These methods are said to have better scalability and other advantages. For example, the simple SlopeOne algorithm is easy to implement, updateable, efficient, expect little from the first visitors and has comparable accuracy to memory-based methods (Lemire & Maclachan, 2005). These properties made it quite successful for web-based recommender systems.

2.2.4 Examples

There are many successful examples of recommender systems on the web, but many of them are running the same algorithms on the same set of items. Resnick's vision of an aggregator-type recommender system (Resnick & Varian, 1997) is not born yet, but there are small separated systems instead.

Application	Data collection	Input data	Recommendation
Youtube	explicit	tags	videos
Google Video	explicit	tags, ratings	videos
Last.fm	implicit	listening	artists, playlists
Pandora	content-based	genes	playlists
iFanzy	implicit	watching, recording	tv shows
IPTV	implicit	watching, recording	tv shows

Table 2: Recommender systems

E-commerce sites, like Amazon, have their own systems kept in secret, but they all depend on item-to-item collaborative filtering based on purchase data. However, there are small startups with innovative ideas for certain domains of application.

Last.fm (http://www.last.fm) is one of the few successful web startup companies outside the US. Their basic purpose is to collect listening habits of the users and build an active community around it. They have built-in support or a plugin for all important media players, so users only have to provide username and password, and the plugin will send all metadata of listened songs to the Last.fm server. It is an easy implicit way to collect data which defines the user profiles. Nowadays, they have a huge database and many interesting services. Last.fm is planning a music video recommendation service, with which users can create a personalized music television.

Pandora (http://www.pandora.com) has a different approach as they are not trying to create a music community. Their goal is to create personalized radios, which play relevant songs for the user. The technology behind Pandora is a result of a tremendous effort and enthusiasm. The project was started by trained musicians in early 2000 to "capture the essence of music at the most fundamental level". They assembled hundreds of musical attributes, called "genes", for The Music Genome Project. In the past 6 years, more than 10000 artists and hundreds of thousands songs were analyzed by music professionals, and they assigned all relevant genes to the songs. With this incredible collection, music similarities can be defined at a gene-level, which makes Pandora a reliable and unique content-based recommender system. This approach needs a great enthusiasm and it's very domain specific. If we consider genes as tags, Pandora is a tag based recommender system.

The recently started Hungarian IP television will have an in-built recom-

mender system, which starts with general social reference groups, but gets more and more personalized by gathering input from the lenght of continous watching, skipping and recording actions.

The iFanzy Electronic Program Guide (EPG) extension is a similar service developed by a Dutch company, which also produces group recommendation and personalized interfaces.

2.3 Other tools

Tagging and recommender systems are quite recent tools for content organization, but there are many more well-established tools that are used for organizing information.

2.3.1 Ratings

Rating is a very basic tool for content organization. Users can assign numerical scores to items representing the subjective quality of the item. On usual websites, the number and average of ratings are presented next to every item, but for example a medium average can mean different things. Maybe it divides its audience because of cultural differences and half of the users rate it great and half of them rate it awful. Medium average can also mean that the item is really of medium quality. Rating habits are extremely different sometimes, so we should take care if we are using aggregated numbers.

Ratings can serve as input data for toplists and recommender systems, so huge amount of ratings are needed in these cases. The Youtube rating system consists of 5 clickable stars, which became very popular so it is used in most video sharing sites and applications.

2.3.2 Toplists

Toplists are basically ordered lists of items according to some parameters. In the case of audiovisual content "most played", "most recent", "top rated" and many other toplists are used. These lists are easy to create and it is a great starting point for new users, so it's widely used on different webpages.

The included items can be filtered by time, category, language, and many more, so with all these dimensions we are able to find the most played Hungarian sport videos last week. For example, Youtube is able to filter by time, category, language and various sorting aspects.

Recent developments include hype lists, which show the top movers of the given period compared to the past. Last.fm uses this list to show the artists with popular new albums or songs.

Toplists are also used in communities to rank users, so top contributors and freeriders can be distinguished.

2.3.3 Flagging

When user-uploaded content is the main content source for a system, it's hard to avoid spam, copyright infringement and the appearance of offensive material. In the past few years, moderators had the role to look through and delete inappropriate items. The number of user-uploaded content is growing so fast that moderators cannot keep up with it. A very simple but useful way is to let users flag items as inappropriate. These items can be reviewed by a moderator or get automatically deleted after a certain number of flags.

This simple technique is used on most social websites. Youtube users can flag items as inappropriate, while Last.fm mainly uses flagging to correct mispelled or false data.

2.3.4 Annotation

Annotation is used for adding metadata to documents, images, videos and webpages. They appear on a layer top of the original content, so it doesn't modify the file, just presents additional information. The most popular tools are sticky notes for documents and webpages, and markers for images and videos. It is a more flexible tool than tagging, as it allows the selection of specific parts from a resource. On the other hand, it expects a lot of work from the user, as he/she has to add the exact position and the description too, which is obviously harder than tagging.

Popular examples include Mojiti (http://www.mojiti.com) and BubblePly (http://www.bubbleply.com) which allows the user to add subtitles, speech bubbles, comments, free writing, and even audio and video commentary. Veotag (http://www.veotag.com) creates multi-level menu of the video content and makes it possible to jump to the most interesting parts in long files. Click.tv (http://www.click.tv) has a similar approach that lets users comment any part of a video.

Site	Ratings	Toplists	Flagging	Annotation
Youtube	yes	yes	yes	no
Flickr	no	no	yes	no
Google Video	yes	yes	yes	no
Last.fm	no	yes	yes	no
Mojiti	no	yes	no	yes
BubblePly	no	no	no	yes

Table 3: Other tools of social websites

2.3.5 Versioning

Versioning means keeping previous versions when a file or metadata is changed. In the programming world it is also called version control, revision control or code management. A version control system can recall previous versions of a file or metadata, and show changes between different versions. The main role of versioning is lowering the entry to open development/content creation: because mistakes are easy to repair and changes are easy to oversee, contributors don't need to be fully trusted, and don't need to be experts from the beginning.

In the case of audiovisual content, versioning also means that remixes and original versions should be connected. When remixes are created inside the system, it can be tracked automatically, but as soon as it leaves and re-enters the system, we have no idea about version connections. In specific communities users add these connections manually (e.g. http://ccmixter.org and http://the-breaks.com/).

2.4 Toolset of p2p applications

2.4.1 Joost

Joost is the next project of Skype and Kazaa founders building on their superpeer technology. It is a system for distributing television shows and other forms of video over the web using p2p television technology.

Joost's aim is to create a p2p television platform with interactive extensions. It has rating functionality, and many tools for online communication like instant messaging, chat and message boards. Currently, in beta state, it has no support for tagging and content recommendation.

Joost is mainly focusing on professional content creatorsand television channels.

2.4.2 Azureus Vuze

Azureus Vuze is the public beta version of the former code-named Zudeo application. It is an extension of the popular Bittorrent client, called Azureus.

It has central channels for professional content, but all registered users can share content too. The uploader can add tags to content items producing a narrow folksonomy. Vuze has many different toplists and users can rate, flag and comment on all items.

The torrent files are stored centrally on Vuze's tracker server, so there is no need for complex distributed algorithms.

2.4.3 Tribler

Tribler is a joint effort of two Dutch universities. Its aim is to create a social-based p2p file sharing application (Pouwelse et al., 2006). Currently, Tribler lacks tagging functionality, but the newest interface plans seem to include tagging and simple tagclouds for content and users, producing a broad folksonomy.

Furthermore, Tribler has one of the first implementation of a distributed recommender system. The algorithm works on implicit binary data, like download history (Pouwelse et al., 2006). It uses the classic Pearson correlation well known from centralized recommender systems. The spreading algorithm is called Buddycast and it's using an epidemic protocol to exchange download histories.

Attribute	Joost	Vuze	Tribler
Folksonomy	no	narrow	not yet
Recommendation	no	no	download-based
Toplists	no	extensive	popularity
Rating	yes	yes	no
Flagging	no	yes	no
Messaging	extensive	comments	no

Table 4: Attributes of p2p applications

The files section has extensive lists of content items that can be ordered and filtered by various aspects. It also provides a popularity toplist which is computed from the number of active peers.

The Tribler system is completely distributed and it's able to import content from web-based services like Youtube. It has other tools like social networking, cooperative downloading and many enhancements compared to other BitTorrent applications, but they don't have a direct effect on content organization and discovery, henceforth they are outside the scope of this paper.

3 New trends and theories

In this chapter I will present some new ideas from recent research papers, which might change the way we think about content organization and discovery.

3.1 Visualizing tags

The tagcloud is a popular method to support navigation and retrieval with visualization, but it can be used to present even more information in a small space than nowadays. Many researchers argue that current tagclouds using inline HTML are wasting too much space and tags are alphabetically ordered, which makes no sense when more expressive orderings are possible (Kaser & Lemire, 2007).

Tagclouds give us a bird-eye view for the resource. They can be used to describe collections of items, groups of people, or even individual items and users when there are a large amount of tags assigned. The proposed enhancements can be divided into two groups: tag weighting and tagcloud layout.

Nowadays, selection and weighting is done by simply the tag's frequency, but it badly characterizes objects with many popular tags. This problem can be partly solved by creating different weighting functions which reduce overlapping of the most popular tags (Hassan-Montero & Herrero-Solana, 2006).

Alphabetically ordered inline HTML tagcloud layouts are the most popular on the internet nowadays. The first problem is that this ordering is only useful when viewing our personal tagcloud, which we obviously know the tags for. In case of browsing a previously unknown tagcloud, which is far more common, it is perceived as a chaotic tag soup (Hassan-Montero & Herrero-Solana, 2006). Different clustering techniques (like k-means) can create tag similarity groups, so these tags can be presented next to each other on the interface. Clustering is also useful for creating a flexible hierarchy for the tags. Second, inline HTML wastes too much valuable space because of the different font sizes in th same row. Newest papers offer great solutions, for example the use of Electronic Design Automation (EDA) and advanced HTML-CSS techniques like nested tables (Kaser & Lemire, 2007).

3.2 New techniques for recommender systems

The importance of recommendation was proved by Netflix, the largest online DVD rental service, in October 2006. They announced a competition to beat their recommender algorithm by 10 percent in predicting user ratings. The grand prize is 1 million dollars, which sounds very high for a seemingly simple problem. This competition is an important milestone in the history of recommender systems, and it marks that these systems will be in heavy use in the future.

As of April 2007, after half year of tremendous research effort from many universities and research institutes, the leading team has reached 7 percent improvement over the reference algorithm. Members of the top 20 are using novel methods for predicting user behavior like neural nets and dimension reduction. In 2007, the ACM Knowledge Discovery and Data Mining conference will have a dedicated workshop for the Netflix Prize, where most of the leading teams will share some knowledge about their methods. In October 2007, a whole ACM conference will be dedicated to recommender systems, so we can expect that brand new methods will be presented this year and recommender systems will improve and become more important in everyday use of the internet.

4 Content organization and discovery in P2P-Fusion

In this section I will try to provide a coherent specification of content organization and discovery tools for P2P-Fusion. It's aim is to combine the best solutions from real applications and recent research ideas, but still support realization in a distributed environment.

To present typical use, we have laid down the channel concept in earlier papers, which means that collections of multimedia files can be organized into playlists. These playlists can be automatically generated or manually set up and administered. Automatic channels are connected to content items, tags, and we can generate automatic channels for users and groups if they wish to. Moreover, users and groups can manually set up channels for their needs.

4.1 Tagging

In earlier papers, we defined 3 different types of tags. Two for objective description and one to express emotional attitude.

Tags These are descriptive keywords for the content.

Flags A pre-defined (but expandable) category of special tags with icons, indicating legal status, or other status in the workflow, like "to be translated" or "to read".

Badges Also a pre-defined (but expandable) category of special tags with icons, expressing emotional attitude, like "funny" or "great".

We believe that users will prefer using flags and badges, because icons describe emotions and status better. Hence, simple tags will be more descriptive and useful for filtering and recommendation.

To have a vivid broad folksonomy, we need to make assigning tags, flags and badges easier and widely used. First of all, for a broad folksonomy, it is important that tags should have different relevancy for the item. For example, a video containing all Premier League goals from last week should be tagged by "football" and "goals" many times, but "Manchester United" is less relevant. Users are unlikely to type in tags that are already assigned by others, so we should let them easily duplicate that tag. We should place a little "plus" icon next to the tags, so the user can easily assign it again.

Attribute	Simple Tagging	Tags,Flags,Badges
Folksonomy	broad or narrow	broad
Descriptiveness	medium	good
Representation	tagcloud	tagcloud, flags, badges
Difficulty of use	easy	easier
Vocabulary	broader	broad

Table 5: Attributes of different types of tagging

Secondly, we should help typing new tags. One of the Tribler research papers has great ideas for this tag suggestion, so it can be used for P2P-Fusion too (Clements et al., 2007).

Objects in broad folksonomies can be visually represented as tagclouds. Tagcloud representation gives a great bird-eye view on the object. We can use the assigned tags for items, or the aggregated tagcloud of the downloaded items for user. We can also create tagclouds on a higher level for groups, channels, or even for the entire mediaspace.

4.2 Tag-based recommender system

If a vivid broad folksonomy is present, we can build advanced services on it. Recommender systems used to work on huge amount of ratings as input data, but if we have descriptive metadata of high quality, we can also use that for the recommendations.

As users watch videos, the most frequent tags describing those videos can be used to describe the users' profile. For example, if the user watched hundreds of videos tagged with "snowboard", this tag can be added to the profile with a calculated weight. The weight should depend on the total number of videos watched by the user, the total number of videos tagged with "snowboard", and the number of videos in the intersection of these two sets. In case of a broad folksonomy, the numbers can be refined by calculating a tag's weight for each video and aggregate it for the user's profile.

The user's preferences can then be described by a tagcloud weighted according to the calculations, but this tagcloud is just a starting point. The user should be able to change the tagcloud and the underlying data by simply deleting or adding tags in the tagcloud. Furthermore, the weights can be refined by little plus and minus icons next to each tag, so a suitable tagcloud

Attribute	Collaborative Filtering	Tag-Based
Input data	rating	tag
Difficulty of use	easier	easy
Runtime	O($item^2$)	O($tag^2 item$)
Ordering	expected rating	expected topic relevancy
Starting time	medium (needs some ratings)	long (needs many tags)

Table 6: Attributes of different types of recommender systems

can be created easily.

The recommendation is easy to do after the weighted tags are ready. We can create vectors from the weights of the user and video items, and take their dot product which is easy to compute. Then the items should be listed in descending order.

Tag-based recommendation is purely based on the folksonomy created by the users, but it can be combined with rating-based methods to create accurate and quality recommendations.

4.3 Comparison table of p2p applications

Attribute	Joost	Vuze	Tribler	Fusion V2
Folksonomy	no	narrow	not yet	broad
Annotation	no	no	no	yes
Recommendation	no	no	download	download & tag
Toplists	no	extensive	popularity	extensive
Rating	yes	yes	no	yes
Flagging	no	yes	no	yes
Messaging	extensive	comments	no	extensive
Access	no	DRM	no	groups
Remixing	no	no	no	yes

Table 7: Attributes of p2p applications

References

[1] Babarczy, E. and Halácsy, P. and Szakadát, I. (2004) *Keresés, relevancia, bizalom, autoritás a hálózaton* in Magyar Távközlés, 2004/2.

[2] Clements, M. and Wang, J. and Yang, J. and de Vries, A. and Reinders, M. (2007) *Personalization of Social Media* at https://www.tribler.org/TagBasedReccommendation

[3] Goldberg, D. and Nichols, D. and Oki, B.M. and Terry, D. (1992) *Using collaborative filtering to weave an information tapestry* in Communications of the ACM, Volume 35, Issue 12, p61-70.

[4] Golder, S.A. and Huberman B.A. (2006) *Usage patterns of collaborative tagging systems* in Journal of Information Science, Volume 32/2

[5] Gomes, L. (2006) *Will All of Us Get Our 15 Minutes On a YouTube Video?* in The Wall Street Journal, 30 August 2006.

[6] Hassan-Montero, Y. and Herrero-Solana, V. (2006) *Improving Tag-Clouds as Visual Information Retrieval Interfaces* in Merida, InSciT2006 conference.

[7] Kaser, O. and Lemire, D. (2007) *Tag-Cloud Drawing: Algorithms for Cloud Visualization* in WWW2007, Banff, Canada, 8-12 May 2007.

[8] Lemire, D. and Maclachlan, A. (2005) *Slope One Predictors for Online Rating-Based Collaborative Filtering* in SIAM Data Mining (SDM05), Newport Beach, California, 21-23 April 2005.

[9] Linden, G. and Smith, B. and York, J. (2003) *Amazon.com recommendations: item-to-item collaborative filtering* in IEEE Internet Computing, Volume 7, Issue 1, p76-80.

[10] Mathes, A. (2004) *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*, at http://blog.namics.com/archives/2005/Folksonomies_Cooperative_Classification.pdf

[11] Pew Internet & American Life Project (2007) *Report on Tagging* at http://www.pewinternet.org/pdfs/PIP_Tagging.pdf

[12] Pouwelse, J. and Garbacki, P. and Wang, J. and Bakker, A. and Yang, J. and Iosup, A. and Epema, D. and Reinders, M. and van Steen, M. and Sips, H. (2006) *Tribler: A social-based peer-to-peer system* in IPTPS'06

[13] Resnick, P. and Iacovou, N. and Suchak, M. and Bergstrom, P. and Riedl, J. (1994) *GroupLens: an open architecture for collaborative filtering of netnews*, ACM Press, New York.

[14] Resnick, P. and Varian, H.R. (1997) *Recommender systems* in Communications of the ACM, Volume 40, Issue 3, p56-58.

[15] Riedl, J. (2006) *Open Issues in Recommender Systems* in The Present and Future of Recommender Systems, Bilbao, Spain, September 12-13, 2006.

[16] Salton, G. and Buckley, C. (1990) *Improving retrieval performance by relevance feedback* in Journal of the American Society for Information Science, Volume 41, Issue 4, p288-297.

[17] Sarwar, B. and Karypis, G. and Konstan, J. and Reidl, J. (2001) *Item-based collaborative filtering recommendation algorithms* in Proceedings of the tenth international conference on World Wide Web, ACM Press, p285-295.

[18] Shardanand, U. and Maes, P. (1995) *Social information filtering: algorithms for automating "word of mouth"* in Proceedings of the SIGCHI conference on Human factors in computing systems 1995, 210-217.

[19] Smith, G. (2004) *Folksonomy: social classification*, at http://atomiq.org/archives/2004/08/folksonomy_social_classification.html

[20] Vander Wal, T. (2005) *Explaining and Showing Broad and Narrow Folksonomies*, at http://www.vanderwal.net/random/entrysel.php?blog=1635

[21] Zollers, A. (2007) *Emerging Motivations for Tagging: Expression, Performance, and Activism* in WWW2007, Banff, Canada, 8-12 May 2007.