PHAROS – Personalizing Users' Experience in Audio-Visual Online Spaces

Ling Chen, Claudiu S. Firan, Wolfgang Nejdl, Raluca Paiu
L3S Research Center
Appelstr. 4
Hannover 30167, Germany
{chen, firan, nejdl, paiu}@L3S.de

ABSTRACT

The large volume of user generated content, although sometimes amateurish, represents a valuable source of information for audiovisual service providers. For example, companies and organizations can efficiently get feedback from consumers observing their online interaction with social media providers. Offering accurately personalized services is possible now, since users provide more personal information about themselves openly, which was previously much more difficult to perceive and measure.

In PHAROS, we aim at exploiting the new and freely available data to improve users' online experience with respect to their interaction with new media. We focus on building technologies, which bridge the gap between the availability of information (both in form of descriptions of content, such as annotations, and user interests and preferences) and the use of it, for augmenting traditional search and retrieval methods or for personalization purposes. In this paper, we describe how this external information can be brought into PHAROS and how it is used to support users, also describing the multiple components supporting this process.

1. INTRODUCTION

The amount of data available on the Web, in organizations and enterprises, is multiplying and data is increasingly becoming audiovisual. Search has become the default way of interacting with content and the ever-increasing data complexity leads to the necessity of a coherent approach to the growing variety of audiovisual formats, standards and tools. Users find themselves overwhelmed by the multitude of new audiovisual search tools, while businesses are at loss for stable direction. The growth of data volume is rapidly shifting to audiovisual content, yet the technologies that allow processing and retrieval of this content are either mainly experimental, or only vaguely capable of handling true queries and content. Audiovisual search is therefore one of the major challenges for organizations and businesses today, and search-based technologies which can provide contextually relevant, integrated and scalable access to distributed and heterogeneous collections of information are essential.

The PHAROS[1] European Integrated Project (Platform for searcHing of Audiovisual Resources across Online Spaces) aims at addressing these challenges and developing an innovative audiovisual technology platform, which takes user and search requirements as key design principles and is deeply integrated with user and context technologies. One of the objectives of the project focuses on the analysis, design and development of context and user technologies taking into account personalization and adaptability. This allows a social audio-visual interaction model to be integrated into the search engine, rather than using a traditional non-participatory information access model. PHAROS creates user interaction models where live user traffic continually improves the user experience via core primitives such as social network analysis or ranking based on trust.

The rest of the paper is organized as follows: In Section 2 we address the progress over the current state-of-the-art in different audiovisual search techniques focusing on user context. We continue in Section 3 with the description of the methodology used in PHAROS for personalization. Then, in Section 4 we present the architectural aspects of the PHAROS platform which support personalization. All modules are described in detail, also presenting the underlying algorithms. We finally conclude the paper in Section 5.

2. PROGRESS OVER STATE-OF-THE-ART

To achieve these ambitious objectives, PHAROS extends the state-of-the-art in the areas of core search technologies, as well as context and user technologies.

Regarding core search technologies, both XML search and content-based search are relevant. Previous work has addressed representation and semantic interoperability [6], as well as XML retrieval [4], [8]. Content based retrieval uses features of multimedia objects to facilitate their retrieval. [2], [10] focused exactly on this topic. However, emerging types of search patterns require both XML and content-based search to be integrated and made mature enough for industrial exploitation. PHAROS extends the state-of-the-art in this area by developing a scalable search platform with advanced query brokering to orchestrate audiovisual information access combining pluggable content-

VLDB '08, August 24-30, 2008, Auckland, New Zealand.

[1] http://www.pharos-audiovisual-search.eu/

based matching engines and schema agnostic XML based search kernels.

Context and user technologies have been tackled from various points of view: social media [9], [7], spam detection and ranking [3], [5] as well as security, trust and privacy [1]. PHAROS addresses all these aspects and more specifically focuses on exploiting user actions and interactions in personal and public spaces to provide advanced and semantically rich recommendations and personalized ranking algorithms. User and community profiles enable extreme precision for search, and exploit all kinds of user-generated metadata. Advanced spam detection algorithms, suitable for personalized ranking, are also provided and new lightweight forms of content protection are investigated.

3. PERSONALIZATION IN PHAROS

PHAROS is placed in a good position with respect to the new Web: there is a lot of momentum in users annotating and tagging audiovisual sources. However, there is a gap between the availability of this information and efficient exploitation for the purposes of improved access to desired content. Further, personalization has been known to suffer from the bootstrapping problem, where the experience of a new user can be unsatisfactory. In addition, there are other avenues of user information where users express their strong and personal opinion. This is the world of blogging – a very popular Web 2.0 phenomenon. Other public spaces including social networking sites and online forums are also rich sources of information about people and their preferences. The vision of PHAROS would be well served by creating technologies to bridge the aforementioned gap. For this purpose, in PHAROS, we currently take into account two types of important social media which is increasingly popular over the Web today: social annotations and weblogs.

Social annotation refers to the user-supplied tags, which are textual labels, to a piece of information on the Web, such as a picture, blog entry, a video clip etc. With the vast development of Web 2.0, social tagging has been a powerful and important feature provided by many social media applications, such as Flickr[2], Del.icio.us[3], and Last.fm[4]. Consequently, large volume of social annotation data can be collected easily, which enables reliable and accurate knowledge discovery. The knowledge embedded in such user-supplied annotation data is believed to be useful in many applications. Therefore, in PHAROS, we also investigate this: in particular, we focus on studying the usage patterns between users and social annotations. We then further explore the usage of discovered patterns in personalized search and recommendations.

Recently, weblogs have become one of the dominant forms of self-publication on the Internet. A weblog, or "blog", is commonly defined as a Web page with a set of dated entries, in reverse chronological order, maintained by its writer via a weblog publishing tool. The contents of entries (posts) are discussions and observations ranging from the mainstream to the very personal. The fast-growing popularity of the blogosphere offers new chances and challenges for Web search. For example, besides searching blogs, we can also analyze weblog communities, as a representative of our target audience, to predict the effectiveness of new recommendations. In PHAROS, by using weblogs we aim at discovering communities, which consist of blog users discussing similar topics in a certain period of time. We then analyze the properties of the identified communities, such as the information diffusion patterns in a community, to create accurate and detailed community profiles. The discovered community information is used to optimize the search and recommendation results for individual users.

4. SOCIAL MEDIA ARCHITECTURE

We start with the description of the architecture of social media modules in PHAROS, to show how the social data is brought to the PHAROS platform and exploited for personalization purposes. There are five modules in total which belong to three layers: ***Offline Analysis***, ***Storage*** and ***Processing***, as shown in Figure 1.

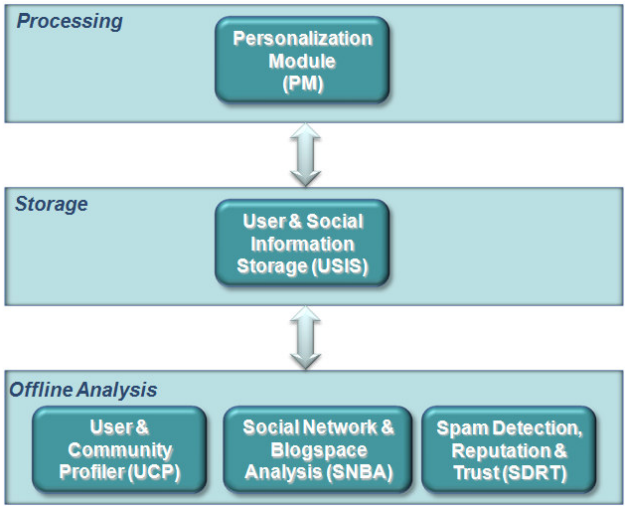

Figure 1. Architecture of Social Media Modules in PHAROS

There are three modules inside the Offline Analysis layer: *User & Community Profiler (UCP)*, *Social Networks & Blogspace Analysis (SNBA)*, and *Spam Detection, Reputation and Trust (SDRT)*. These modules retrieve related social metadata either from the PHAROS platform or from some other sources available on the Web. They further process the collected raw data to extract useful knowledge for other functionalities in PHAROS. In particular, the UCP module focuses on collecting and creating complete and accurate user profiles, as well as community profiles so that precise and personalized search and recommendation can be provided based on these profiles. The SNBA module aims at retrieving social network data, such as friendship network and blogspace information, and analyzing the social network both from a micro-perspective (e.g., the network of a user community) as well as a macro-perspective (e.g., the network of all PHAROS users). Due to the fact that current social media technologies are highly vulnerable to malicious users motivated by both private and commercial interests, the SDRT module is developed to improve the robustness of the PHAROS platform by detecting spam and assigning reputations and trust values to the users involved in social interactions. As SDRT is not implemented yet

[2] http://www.flickr.com/

[3] http://del.icio.us/

[4] http://www.last.fm/

inside the PHAROS platform we will focus only on UCP and SNBA as Offline Analysis modules.

The useful knowledge extracted from social metadata by all of the Offline Analysis modules is recorded in the same storage, the *User & Social Information Storage (USIS)*. Periodic updates are initiated in order to ensure that the stored knowledge is not obsolete. Beside the interfaces with Offline Analysis modules, USIS also interacts with the *Personalization Module (PM)*, located inside the Processing layer, to provide requested information for search and recommendation. The PM manifests the usefulness of social metadata in PHAROS. Various personalization strategies based on different extracted knowledge are developed to finally optimize the search results provided by PHAROS.

The details of each of the modules composing the Social Media Architecture will be described in depth in the following sections.

4.1 USIS – User & Social Information Storage

The *User & Social Information Storage (USIS)* plays a central role inside the architecture of the PHAROS platform, as this is the place where all user-related information is stored.

The functionality of the USIS is divided into the following roles:

- Metadata storage for PHAROS content objects;
- User related data storage and processing;
- External social interaction data storage.

Metadata Storage. Each PHAROS content object can have different types of metadata attached (e.g., tags, comments, ratings, and favorites). Each user having an account in the platform will be able to enrich content objects with metadata. This metadata can be viewed and searched by the users.

User Related Data. All the information stored in the USIS is meant to be later extracted and included into the personalization process. User and community profiling information, in form of both preferences as well as interaction with the PHAROS platform, are stored in here. Several parts build up the user profiles and are stored inside the USIS (details are presented in Section 4.2).

External Social Data. Data from external (not residing inside the PHAROS platform) sources like social networks or collaborative tagging Web sites can be also stored inside the USIS. This data can be used by any of the Offline Analysis modules in order to extract additional data relevant for PHAROS users or content objects.

4.1.1 Internal Structure of USIS

USIS is mainly a storage component, a database, providing several services depending on different data to be stored or requested. Currently, four major storage components reside inside the USIS.

Blog Analysis Storage. Blogs gathered from different sources are stored here. These blogs are further processed, and analyzed in detail by SNBA extracting interesting features and statistics needed by other components (e.g. UCP). The results of the analysis are stored in the Blog Analysis Storage as well.

Interaction Logs. User actions related to querying, receiving recommendations, and result handling is monitored by the PHAROS platform and are stored at the end of a user session in the USIS. Information stored here includes: what query was entered by the user, what results were viewed, what results were clicked, if the visualization of the results was interrupted prematurely (e.g. a video was not viewed until the end), etc.

User Generated Metadata. All user generated metadata resides in this storage component. This includes tags, comments, ratings and favorites. Information is stored as to which user added what metadata to which content object. In this way the metadata is filtered and statistics are computed regarding a user, a piece of metadata (e.g. some specific tag), or a content object.

User Profiles & Groups. User profiles, user-user relationships, and user-group memberships reside here. User profiles are constructed initially from the personal data a user enters when s/he creates his/her profile. UCP adds more data to the profile as the user starts using the platform.

4.2 UCP – User & Community Profiler

In order to achieve extreme precision in ranking and recommending multimedia content, adaptation of the core technology to user preferences and specific user contexts is necessary. Since most users may be unwilling to explicitly fill in and maintain a personal profile or they might not be able to specify an accurate profile, automatically inferring interests is important. Moreover, for providing high quality personalized services, profiles must be kept up-to-date as interests may change over time. Accurate user profiles often also depend on the community[5] a user belongs to. Therefore, inferring user profiles has to be complemented with the construction of community profiles. The *User & Community Profiler (UCP)* component is in charge of modeling user preferences from both inside and outside the PHAROS platform by collecting and automatically inferring information about users and communities they belong to. To overcome problems associated with modeling and finding communities adequately, it builds upon a model of user and community actions and interactions in social networks (provided by SNBA).

This module takes advantage of the explicit profile information freely provided by the PHAROS users and at the same time extends it with publicly available profile data from the services indicated by the user. Moreover, interests or preferences are implicitly found in concrete user (inter)actions and can thus be modeled from logging user interactions and group behavior within the PHAROS platform. Given this diverse amount of (raw) data about users and communities, the challenge and main focus of research activities within this module is to develop advanced techniques for building user and community profiles detailed, recent, accurate and reliable enough to meet the challenges of precise personalized and context-specific retrieval and recommendation.

In detail, the user or community profiles comprise:

- Explicit user information, given in the account/my profile section in the user interface, including basic data about users (gender, age, language) as well as some general interests;

[5] We use the term "community" when referring to any external social network of people e.g. build on platforms like Flickr; Social Group is used when we refer to the groups that users are actually building within the PHAROS platform

- User generated metadata (tags, comments or favorites) made within the PHAROS platform. Metadata in external Web 2.0 platforms like Last.fm, Del.icio.us or Flickr are analyzed and aggregated to infer preferences;
- Interests of individual user and communities extracted from the blogspace;
- Usage history. Issued queries and click through data from the interaction logs are used to show recently accessed resources ("charts") and serve as implicit feedback to infer likes and dislikes;
- Friendship or contact relations between users both inform about similarities or common interests and may be used for restricting recommendations based on privacy concerns;
- Similar users (neighbors) are automatically detected by combining data about users as described above.

Since users may have different preferences with respect to different situations, users can have multiple user profiles comprising the attributes listed – one for each of their various contexts (like "work", "leisure" etc.). For effectively supporting distinct user profiles, current active contexts have to be identified accurately to add the information to the right place. However, a default profile giving all information available is supported. To take into account most recent user and community data, profile updates are scheduled according to availability of new data.

4.2.1 Functionalities of UCP

At the current stage, the UCP module performs the analysis as follows.

Log Analysis. What resources a user searches for, which multimedia resources he actually accesses (for how long), whether he even recommends them to other people, as well as which persons he frequently interacts with, provide a lot of valuable data about a user's topics or preferences and probably typical behavioral patterns. In contrast to explicitly provided profile information, such implicit feedback to resources and people has to be analyzed to infer meaningful and generalizable user attributes to optimize personalization. For example, the user evaluation of the result set (skipped and clicked items) helps to infer new associated terms by getting keywords from resources implicitly judged as relevant. On the other hand, similar resources listened to or watched can be exploited to find similar queries or even super-ordinate topics. In general, just building the list of (recently) seen items - usage history of a user – alone is very useful for profiling interests with respect to finding similar users or similar resources. Interactions within groups and with friends or unknown people are analyzed to model the social network of a user which can again be used to infer commonalities and preferences. Other patterns to be mined from action sequences (like system internal navigation paths) help personalizing view settings and creating navigational short cuts as well as to inform about general usability issues to be improved. All these extracted and inferred information are translated into specific attributes and written to the user profile.

Annotation Analysis. By adding tags, comments or favorites to multimedia resources seen within PHAROS, users tell us (implicitly) about what they like, don't like or what topics they are interested in, as they organize their resources around it. Therefore, Annotation Analysis gathers this kind of data and analyzes it in depth to make it available for profiling and search.

Opinion Mining. For comments, stopword removal and term normalization are standard procedures, important keyword extraction and Sentiment Analysis / Opinion Mining are more advanced techniques. This aims at analyzing any free textual annotations about content objects within PHAROS or elsewhere on the Internet. From these textual annotations new tastes about audiovisual objects are deduced, and the user profile is updated with this new knowledge.

Tag Analysis. Tag analysis comprises first of all normalization and the inclusion of alternative, synonymous labels. Also absolute and relative frequency information is calculated for individual tags as well as co-occurrence relationships that may help to dissolve term ambiguities or to refine queries by synonyms / strongly associated terms. To exploit the potentially huge effort a user already invested in tagging interesting Web pages, songs or pictures in one or the other Web 2.0 platforms, tag analysis also fetches and analyzes the user's tags from external sites like Last.fm and Del.icio.us.

Profile Building. Finally, the results of the above mentioned single analyses have to be merged and enriched with the information explicitly given by the user. Both types of information go directly into the profile under the corresponding attributes. Note that this may mean merging or resolving conflicts about preferred topics identified in the single analyses.

4.3 SNBA – Social Networks & Blogspace Analysis

The *Social Networks & Blogspace Analysis (SNBA)* module aims at gathering and analyzing social network data coming from blogs or friendship networks for discovering additional knowledge which can be applied to improve the search and recommendation results in the PHAROS platform.

There are many different types of blogs, differing not only in the type of content, but also in the way that content is delivered or written. However, for our analysis we focus on personal blogs, as this type of blogs – on-going diaries or commentaries by individuals – reveal the most personal information about their authors. Since all blogs are on the internet by definition, they may be seen as interconnected and socially networked. Several features permit bloggers to link to each other's blog pages: the so-called "blogrolls" lists one's favorite blog list in a frame inside their own blog page. These links represent other authors' blog pages that this author considers interesting and frequently visits for reading and / or directly commenting. In a sense this feature is similar to the in-links of a Web page: they inject some importance to the blog pages they target by the fact that the author of the blog page lists these links on his own and indirectly shows that there are some trusted blog sources, worth reading. Besides, the blogrolling phenomenon is somewhat reciprocal. By linking to a blog, users are increasing their blog's chances of being linked-to by other weblogs. These links between Weblogs are the "currency" of the Weblog community. The more links one has pointing to his weblog, the more likely he gets a growing audience and high rankings in search engines. A blogroll helps a user get started earning links from other weblogs by expressing affiliations. Permalinks are also a possibility to create social links among bloggers. Unlike blogrolls which point to a blog page, a permalink represents a link to a particular blog post inside a blog page (created by the author of this page) and allows other bloggers to

use it to jump directly to this blog entry. Given the highly dynamic content change of the blog pages, this feature is extremely useful if users want to re-read some very interesting post, which has already passed from the front page to the archives.

Given the aforementioned characteristics of the weblogs and of the blogosphere, it can be easily seen that blogging is inherently a social process, one in which information is created and diffuses (or flows), between bloggers due to bloggers influencing and being influenced by other bloggers. Consequently, the overall objectives of the SNBA module are to:

- Determine ways to capture what is diffusing;
- Determine paths for who influences whom;
- Measure the extent of this influence;
- Exploit this knowledge for personalized ranking and search.

We touch each of the objectives presented above, describing the components and algorithms for exploiting information diffusion for personalized ranking and search.

4.3.1 Functionalities of SNBA

Several components build up the SNBA module, such as Blog Identification and Blog Ingestion etc. We focus on describing the main component Blog Analysis Processing.

The Blog Analysis Processing component is further divided into different sub-modules (see Figure 2). It works as a three-stage process: at the lowest level, a static snapshot representation of the domain is obtained through the Text Mining Module. Based on this, higher level abstractions are built (Topic / Community Detection). Then, dynamics and evolution are handled within the Information Diffusion Module. Finally, a Profile Extraction component aggregate discovered knowledge into users' as well as communities' profiles.

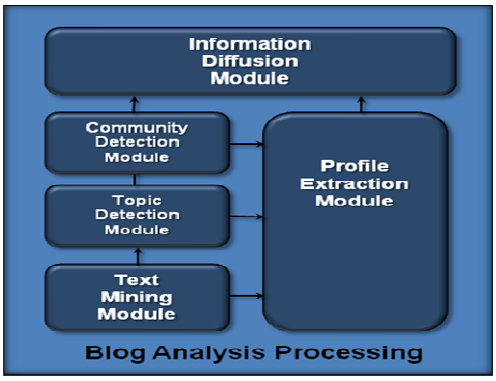

Figure 2. Blog Analysis Processing

The Text Mining Module uses pre-processed blog posts as input for the mining process. The output (typically described by a set a probability word pairs) is used as input for determining the topic of blog posts or of the blogs themselves. Community structures are then created from these underlying descriptions. Topics from blogs and posts written by a user are taken as representations for the user's and her communities' profiles. Once communities have been detected – the dynamics and evolution of information between individuals and communities are mined, in the Information Diffusion Module. Finally, the Profile Extraction Module updates the created profiles with topic, time and information diffusion information etc. Due the space constraints, we ignore the discussion of the algorithms employed by each module. Interested readers can refer to [21] for the detail.

4.4 PM – Personalization Module

The *Personalization Module (PM)* focuses on providing personalized search and recommendation functionality. This component is especially important because it unlocks the value of personal information stored in USIS in order to improve the users' experiences inside the PHAROS platform. PM takes requests from the user and uses information stored in USIS as basis for performing personalization. The relevant information about user interests is computed offline in the UCP and SNBA modules and is then retrieved by PM both during the pre-computation of personalized ranking values, as well as during the model building phase of the recommendation engine. This model is later used to compute recommendations online.

4.4.1 Internal Structure of PM

The component architecture has been designed in order to support *pluggable* recommendation algorithms (see Figure 3), which can be further developed and extended, for example to adapt the behavior of the PM module depending on the context or the user data available, and to try to complement some weaknesses and strengths of the algorithms themselves, by creating hybrid models that combine them.

In case of the Personalized Search Component there are two critical factors for the effective personalization: the quality of the user profile and the query processing time. The user profiles are pre-computed by UCP and SNBA components and stored in the USIS module. The PM module communicates with USIS to fetch and transform these profiles into a format required by different personalization algorithms. During query time, the system sends a request with the original query to PM and receives back a new query which includes the necessary modifications for a better ranking.

One part of the PM module is the *Query Personalization Component* (see Figure 3). When a query comes from the system via the provided API, the *Query Parser* transforms it into internal format for further processing. The *Personalization Selector* component chooses the requested personalization method and asks the *Profile Retriever* for a necessary user or community profile information. The *Query Personalizer* transforms the original query and sends the resulted personalized query back to the system for retrieval.

The *Recommender System Component* also depends directly on the user profiles pre-computed by UCP and SNBA. Once these profiles are retrieved from USIS, the *Modeler* sub-component builds a model of the user preferences. The computation is done offline periodically and the results are also stored back into the USIS. Once the model has been built, the Recommendation Engine is ready to compute the necessary list of personalized recommendations for a given user, as well as her neighborhood (User-based recommendations). Depending on the context, the Recommendation Engine is also capable to recommend similar items given a resource (Item-based recommendations).

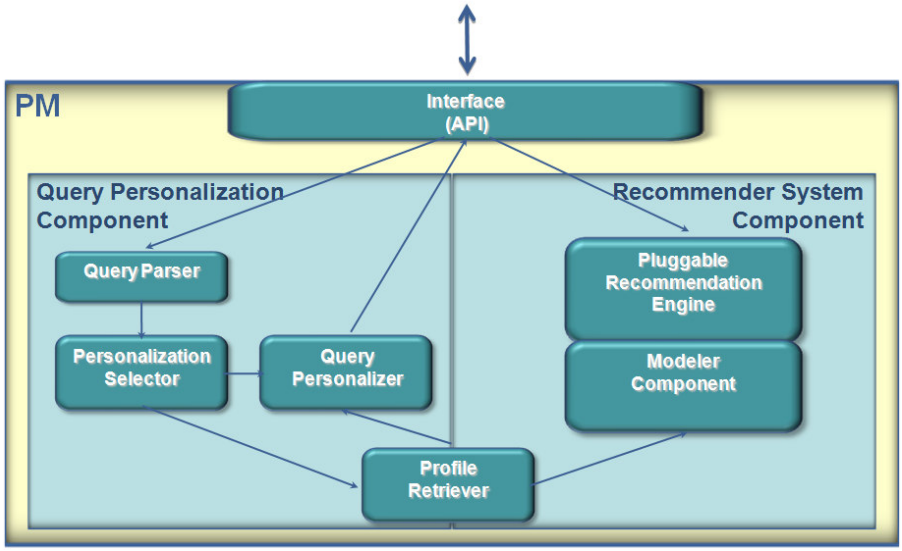

Figure 3. Internal Structure of the Personalization Module

The *Query Personalization* and *Recommender System* are the fundamental components of PM and below we describe each of them in detail.

4.4.2 Query Personalization Component

In addition to general search capabilities, PHAROS provides also personalized search results matching the profile of the user or any groups the user belongs to. Personalization involves both a filtering and ranking of results. Result filtering is used to limit the result set to content, which fits the user's information need, and content the user has permission to view. Ranking of results takes into account user and group preferences and ranks content, believed to be of high relevance to the user, higher than content which is of general interest. Ranking parameters are part of the query, and are inserted by the PM module. Different personalization techniques are developed within Information Retrieval field, like query re-weighting and query expansion, just to name a few. We provide details on implemented methods in Section 4.4.3.

The personalized search capabilities assume re-ranking of the relevant multimedia items using information about previous user's interactions with the PHAROS platform. All historic usage data, such as user's queries, clicked results, tags in use, etc., should be exploited to provide a more precise search output. For providing high quality personalized services, user profiles must be kept up-to-date as interests may change over time.

4.4.3 Personalized Search Algorithms

For the *Query Personalization Component*, we implemented 6 Information Retrieval algorithms for Relevance Feedback and Query Expansion. The algorithms can be used as standalone methods as well as in combination with each other. The effectiveness of the proposed methods has been proved in general text and multimedia retrieval, while their practical usefulness depends on available data and quality of the user profiles:

(1) Fields Reweighting. Results are initially ranked using default values for the given query keywords – fields. Based on previously collected information regarding user tags and associated ratings an algorithm can specify a different weight for each field (query keyword) and this information is then used for ranking. The frequently used user's tags and query fields receive higher weights and results are biased towards them. This relevance feedback technique is based on a well-known Rocchio method [11]. Long-term feedback is obtained from tag usage of the user or user group and is captured from UCP module in the form of a user profile, which tracks user-specific weights and other feedback-based parameters. The vector-space representation of the query is modified so that more important term dimensions are emphasized and similarity between query and each item of interest is affected. Top-N most similar items are then presented to the user.

(2) Results Filtering. Create restrictions based on the user profile, like removing from the ranked list of results the items that the user dislikes. We use Generalized Query Point Movement method [12], where previously received poor ratings from a user are used to extract tags representing what the user does not like. The result items containing such tags are moved down the ranking. The algorithm has a similar mechanism to Fields Reweighting technique, but negative assessments are used to compute the user profiles. This personalization technique is effective when users explicitly mark items as uninteresting.

(3) Query Expansion. Based on the users' tag usage patterns we compute tags' similarity to each other. Additional keywords are added to the query based on preferences from the user profile. This method [13] is one the most frequently used for personalization and can significantly increase recall in situations, where original query does not have enough results. Query expansion is essentially adding new features to the query vector and re-ranking the results accordingly. The initial query terms are still of higher importance for the ranking. The precise values for the algorithm tuning have to be defined based on available data, which can be done as soon as user profiles and interaction histories are collected.

(4) Query by Example. For the scenarios when a user wants to find items similar to a current one we consider tags similarity. A tag description of the selected example is used as a query and a set

of results, excluding original item, is returned to a user. The personalization part of this method assumes additional query modification, based on recent user queries and tags used. A method ranks higher the results which are not only similar to a given example, but also remind previously seen items.

(5) Community-Tag Rank. This algorithm is inspired by work on Topic-Sensitive PageRank [14] and [15], but based on textual similarity rather than link-based similarity. Document model includes all the fields that can be extracted for multimedia content. We create a community-tag vector for each of the identified communities. To make computation scalable we assume that number of communities is significantly lower than a total number of users. A Community-Tag Rank represents a similarity between an item and a community vector, which is composed from tag usage statistics of all users belonging to the community. Communities can be defined based on explicit membership in particular communities and automatically computed clusters of users. A user belongs to one or more communities (topic groups) and we can compute a linear combination of the community vectors for items before query time, which is called Community-Tag Rank. A single score of Community-Tag Rank is associated with each multimedia item—community pair. During the computation of the item-query similarity the personalized Object Rank vector is used as a factor for ranking as a query-independent parameter.

(6) Community-Rating Rank. This method is similar to Community-Tag Rank, but is based on a collaborative filtering rather than document model. A community profile for a Community-Rating Rank computation consists of previously issued users' ratings and independent of multimedia item tags. This average rating allows re-ranking retrieved items with respect to their overall popularity among community members and quality of each returned result. As with Community-Tag Rank, this value is query independent and it is pre-computed. During query time this value is added to the item relevance score.

4.4.4 Recommender System Component

Recommender Systems support people by identifying products or services they appreciate, helping them to face the information explosion, where the complexity of offers exceeds the user's capability to survey them and reach an optimal decision.

Different approaches have been suggested for supplying meaningful recommendations to users and some of them implemented and deployed successfully over e-commerce and services sites like Amazon[6], Netflix[7], or MyStrands[8].

State-of-the-art Recommender Systems mostly use a variant of Collaborative Filtering (CF), an approach to solve the recommendation task that relies on historical data gathered from users, rather than using the information about content. The underlying assumption of the CF approach is that those who agreed in the past tend to agree again in the future, capturing human behavior: people searching for an interesting item of which they have little or no information, tend to rely on friends to recommend items they tried and liked.

[6] http://www.amazon.com/

[7] http://www.netflix.com/

[8] http://www.mystrands.com/

The goal of the PM's *Recommender System Component* is to identify neighborhoods of users with similar taste, based on the profiles built by the UCP and SNBA modules and stored in the USIS. To build a user's neighborhood, the Recommender System Component relies on information of past user interactions (e.g., explicit ratings, tags assigned) or implicit grading methods based on user behavior actions, such as the time spent on a particular item Web page. In order to provide recommendations for a given user, the system uses her corresponding neighborhood to compute a list of items interesting for her. A similar approach is also taken to consider neighborhoods of similar items to be exploited in order to provide recommendations of similar contents and resources.

4.4.5 Recommendation Algorithms

In the case of the *Recommender System Component* the following algorithms are provided, and are also combined with each other to provide the target functionality:

(1) Tag-aware Collaborative Filtering. It exploits the tag-based profiles, in both dimensions (user, tag) and (item, tag) to build user and item neighborhoods in order to compute personalized recommendations [16]. Tag-based user profiles are defined as collections of tags together with corresponding scores representing the user's interest in each of these tags. Once the profiles have been computed, they are arranged in a User-Tag matrix structure, which is then used to derive the recommendations applying CF techniques that group similar users in order to suggest them valuable items that in turn have been inferred by their associated tags.

(2) Standard User-based Collaborative Filtering. It supports (1) and can also be used alone or as part of other Recommendation Engine to exploit different kinds of profiles, and not only explicit ratings as in traditional CF [17], [18]. The recommendations for each individual user are obtained by identifying a neighborhood of similar users and recommending items that this group of users found interesting. The design recommendations described by [19] have been also considered in the implementation of this algorithm.

(3) Standard Item-based Collaborative Filtering. The recommendation task in this case is focused on the items' similarity, rather than on the users' similarity. It also supports (1) and its main objective is to produce a list of recommendations given a target item [20]. This recommendation algorithm uses the item-to-item similarities to compute the relations between the different items. It builds a model that captures these relations and then applies this model to derive the top-N recommendations for an active user. The model, which at the core is an item-item matrix representation, is built based on the original user-item matrix of user profiles that reflects their aggregated historical information of consumed items. Each item is associated with a vector in the users' space, and these vectors are then used to compute the similarity among the items. Once the similarities have been computed, for each item, just the most similar k items are kept on the model, where k is an input for the algorithm. The model computed is used during the recommendation step, where the goal is to recommend similar items for a given one.

5. CONCLUSIONS & FUTURE WORK

In this paper, we focus on describing social media metadata based personalization supported in PHAROS. The vision for PHAROS

has an important place for user generated social metadata; the use of personalization and recommendation is vital to PHAROS for enhancing the user experience. The content based search service is served better by an accurate and efficient social based search service. Consequently, social media data analysis and processing plays an important role in audiovisual online spaces.

Particularly, we describe four social media related modules developed in this project. Two analysis modules, User & Community Profiler (UCP) and Social Networks & Blogspace Analysis (SNBA), aim at performing analytic study on various user-generated social media data and extracting knowledge relevant to users' interests. This extracted knowledge is further exploited by the processing module, Personalization Module (PM), to enhance users' personal search experience. A data storage module, User & Social Information Storage (USIS), is also provided to accommodate not only raw social media data but also processed and extracted information from the raw data.

There are still a few open issues in successful exploring social media data for personalization within PHAROS, such as scalability and robustness. Our ongoing work include improving the algorithms employed by each module to address these issues, as well as conducting more research and developing work in optimizing the functionalities provided by social media modules by large scale.

6. ACKNOWLEDGMENTS

This work was partially supported by the PHAROS project funded by the European Commission under the 6th Framework Programme (IST Contract No. 045035).

7. REFERENCES

[1] Aichroth P., Puchta S., Hasselbach J. Personalized Previews: An Alternative Concept of Virtual Goods Marketing, *Virtual Goods Conference, 2004.*

[2] Aslam J., Montague M. Models for Metasearch. *SIGIR, 2001.*

[3] Bharat K., Henzinger M. R. Improved algorithms for topic distillation in a hyperlinked environment, *SIGIR, 1998.*

[4] Carmel D., Maarek Y. S., Mandelbrod M., Mass Y., Soffer A. Searching XML Documents via XML Fragments. *SIGIR 2003.*

[5] Carvalho A., Chirita P. A., Silva de Moura E., Calado P., Nejdl W. Site Level Noise Removal for Search Engines. *WWW, 2006.*

[6] Dong X., Halevy A. Malleable Schemas. *WebDB, 2005.*

[7] Ghita S., Nejdl W., Paiu R. Semantically Rich Recommendations in Social Networks for Sharing, Exchanging and Ranking Semantic Context. *ISWC 2005.*

[8] Kakade V., Raghavan P. Encoding XML in Vector Spaces. *ECIR, 2005.*

[9] Kumar R., Novak J., Raghavan P., Tomkins A. On the bursty evolution of blogspace. *WWW, 2003.*

[10] McDonald K., Smeaton A. A Comparison of Score, Rank and Probability-based Fusion Methods for Video Shot Retrieval. *CIVR, 2005.*

[11] Rocchio, J. J. Relevance feedback in information retrieval. *Prentice-Hall, 1971.*

[12] Ortega-Binderberger, M., & Mehrotra, S. Relevance Feedback in Multimedia Databases. *In Handbook of Video Databases: Design and Applications, 2003.*

[13] Qiu, Y., & Frei, H. Concept Based Query Expansion. *SIGIR, Pittsburgh, 1993.*

[14] Haveliwala, T. H. Topic-sensitive PageRank. *Proceedings of the Eleventh International World Wide Web Conference, Honolulu, 2002.*

[15] Chirita, P. A., Nejdl, W., Paiu, R., & Kohlschütter, C. Using ODP Metadata to Personalize Search. *Proceedings of the 28th ACM International SIGIR Conference on Research and Development in Information Retrieval, Salvador, 2005.*

[16] Firan, C., Nejdl, W., & Paiu, R. The Benefit of Using Tag-Based Profiles. *Proceedings of the 2007 Latin American Web Conference, Santiago, 2007.*

[17] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proceedings of Conference on Computer Supported Coorperative Work , 1994*

[18] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM, 1997.*

[19] Herlocker, J. L., Konstan, J. A., & Riedl, J. An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms. *Information Retrieval, 2002.*

[20] Deshpande, M., & Karypis, G. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems, 2004.*

[21] A.Stewart, L. Chen, R. Paiu & W. Nejdl. Discovering Information Diffusion Paths from Blogosphere for Online Advertising. *ADKDD, California, 2007.*