Computers in Human Behavior xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers in Human Behavior

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

A user-centered and group-based approach for social data filtering and sharing

Xuan Truong Vu *, Marie-Hélène Abel, Pierre Morizet-Mahoudeaux

UMR CNRS 7253 Heudiasyc, Unviersity of Technology of Compiegne, France

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Social networking sites
Social Data Aggregation
Information filtering
Information sharing
Content management
Collaboration

ABSTRACT

Social networking sites (SNSs) like Facebook, Google+, Twitter, LinkedIn have become a very important part of our daily life. People are connected to multiple SNSs for networking, communicating, collaborating, sharing and seeking for information. Although, the diversity of current SNSs increases and enriches our online experience, they cause some problems. One of the major issues is that users are often overwhelmed by the huge number of social data. It is even worse as these social data are scattered across disconnected SNSs. To address such problems, we propose a user-centered and group-based approach for social data filtering and sharing. First, it allows users to aggregate their social data from different SNSs and to extract relevant contents. Users explicitly define their interests via specific queries, using information filtering techniques, the system will retrieve new corresponding contents. Second, it is expected to extend its first user-centered purpose by allowing group-based information sharing and management. Users can share some part of their own social data with and collectively define the information organization within their respective groups. To describe further and illustrate our proposed approach, a system architecture and a prototype are also presented in this paper. A primary test was carried out and showed encouraging results confirming the added values of our approach.

1. Introduction

With social media sites, social networking sites (SNSs) have made up today social Web sites (Kim, Jeong, & Lee, 2010) and become a very important part of our daily life. People spend more time in SNSs than ever for three main uses: (i) gathering, sharing information and contents, (ii) keeping in touch with family, friends, colleagues, and (iii) finding, making new friends (Bonds-Raacke & Raacke, 2010).

There are a large number of SNSs available (Solis, 2013). Facebook, Google+, Twitter and LinkedIn are some of the most successful examples. By attracting millions of active users around the world, they occupy a central place in the social media landscape (Cavazza, 2014). Each of them provides users with its different and unique features: Facebook allows third party applications to build on its application programming interfaces (APIs), Google+ embeds other Google's services, Twitter offers real-time micro-blogging, and LinkedIn focuses on professional networking. Thus, it is very common that one user is simultaneously connected to several SNSs.

Despite such growing popularity and a wide range of provided services, current SNSs raise some issues associated with users' social lifestyle and interaction such as privacy, identity theft, addiction, and spread of bad information (Kwak & Lee, 2011). While these issues are very important in SNSs, they are however not in the scope of our work. We are more concerned with issues related to the usage experience of SNSs, two of which have been particularly the focus of our work: (i) *Information Overload*, and (ii) *"Walled garden" Problem*.

Information Overload: Users are increasingly facing with information overload in SNSs (Borgs, Chayes, Karrer, & Meeder, 2010). They are often overwhelmed by the huge number of coming information. For example, a normal user has, on average 338 friends on Facebook (Smith, 2014) and 208 followers on Twitter (Smith, 2014). As such, he/she receives per day hundreds of diverse *social data* which can be posts, status, updates, photos, videos, links, tags, check-ins, etc., via these social connections. This is much beyond what users could process manually. Moreover, most recent SNSs put all social data generated by a user's social connections into a single stream (e.g. Facebook New Feed, Twitter User Timeline) and sort them in a chronological order. New social data are constantly appearing within and flooding these streams. As a result, many important and interesting pieces of information remain unnoticed by the user, whereas lots of irrelevant and not worth reading contents keep showing up.

"Walled Garden" Problem: Current SNSs, namely Facebook, Google+, Twitter, Linkedin, all operate as "walled gardens", where user data and generated contents are exclusive to the SNSs Chisari

* Corresponding author.
E-mail address: xuan.vu@hds.utc.fr (X.T. Vu).

http://dx.doi.org/10.1016/j.chb.2014.11.079
0747-5632/© 2014 Elsevier Ltd. All rights reserved.

(2009). Most importantly, they provide very little interaction with each other. Users are always required to create a new profile when joining a SNS. Then, they have to manage their social data as well as to keep track of all recent updates across SNSs. On the one hand, it worsens the problem of information overload as social data are scattered over different SNSs. On the other hand, the users of a given SNS are not able to interact and share interesting information or useful contents with the users of another SNS.

To address both the problems, we propose a user-centered and group-based approach for social data filtering and sharing. First, this approach enables the aggregation of social data from different SNSs and the extraction of relevant contents. The users' social data are constantly retrieved and gathered from different SNSs by means of dedicated programs previously granted by the users. Every new social data is then processed, enriched and indexed before being filtered and selected according to the users' interests, which have been explicitly defined beforehand by the users by using specific queries. The users can therefore quickly access and view the contents matching a particular topic by selecting it.

Beyond such a user-centered objective, the approach is expected to also allow group-based information sharing and management. To be more practical, the users can join one or several groups driven by common interests or topics where they can share some part of their social data. Conversely, the users can reach more appropriate contents. In particular, each member of a given group is encouraged to contribute their own intelligence to collectively define and improve the group's evolving areas of interest.

Besides these two principal contributions, we also present, in this paper, an extensible system architecture for implementing our approach. The system is composed of a number of specific modules, which are easy to improve and extend.

The rest of the paper is organized as follows: In the next section, we present a summary of related work on Social Network Aggregation, information filtering, and content curation and sharing. Then, we introduce our user-centered and group-based approach for social data filtering and sharing. After, we describe the important modules for the implementation of the system. In Section 5, we present our web-based prototype and discuss some encouraging results of the first test using this prototype. Finally, we summarize the contributions of our work.

2. Related work

Our work attempts to implement a new emergent paradigm called *Social Internetworking System* (SIS), where a SNS can be seen as a part of a more complex system comprising many users, social networks and resources (Meo, Nocera, Terracina, Ursino, & De Meo, 2011). SIS has raised more and more interests from researchers for enabling strategic applications whose main strength is the integration of different communities that nevertheless preserves their diversity and autonomy (Buccafurri, Lax, Nocera, & Ursino). While the Social Network Analysis can play a very important role in the SIS scenario (Fan & Gordon, 2014), we are more interested in information management approaches that can fit both personalized and collective uses. The approach presented in this paper involves three different fields such as *Social Data Aggregation*, *information filtering*, and *content curation and sharing*. We review below some representative works related to these fields. Then, we discuss about their respective limitations, and outline the added values of our proposed approach.

2.1. Social Network Aggregation

Social Network Aggregation is the process of collecting, aggregating and organizing data spread across multiple SNSs. Such process could be brand-oriented as well as user-oriented. The former allows brands to track and capture as much as possible public messages and comments concerning their reputation (Fire, Puzis, & Elovici, 2013; Gao, Wang, Luan, & Chua, 2014), while the latter attempts to organize a user's social networking experience as a whole.

Here, we solely discuss about the user-oriented aggregation, the first challenge of which is to identify unique users across SNSs. *People search engines* such as Peekyou,[1] Pipl[2] allow to search the different social profiles of a user based on public personal attributes (e.g. name, username, email or location). However, a user may set different values to these attributes, even leave them undefined that makes user identification incomplete. Google proposed an alternative, named *Social Graph API*,[3] which crawls users' personal web pages (e.g. blogs) and extracts links referring to their social profiles. Unfortunately, it is no longer available. In this work, we do not deal with this non-trivial problem as we directly ask users for indicating their different social identifiers.

Another important requirement for integrating social data from different SNSs is to define a common model for the representation of social data. A number of light-weight ontologies have therefore been developed. Some of them such as Friend Of A Friend (FOAF),[4] Semantically-Interlinked Online Community (SIOC),[5] Weighted Interests (WI),[6] Open Provenance Model (OPM)[7] and Activity Streams[8] have already been widely adopted. The authors of Kapanipathi and Orlandi (2011) and Orlandi, Breslin, and Passant (2012) showed that these vocabularies could be combined and used as a domain ontology for integrating social data from many SNSs. Furthermore, they proposed to use linked open data (e.g. DBpedia,[9] OpenCalais[10]) in order to semantically enrich aggregated data. Although, it could be very interesting to follow such a Semantic Web principle, we describe, in the next section, another generic model which fits better our previously cited objectives.

Commercial solutions such as FriendFeed,[11] Hootsuite,[12] Flock[13] attempt to implement Social Network Aggregation, and are so called *Social Network Aggregator* (SNA). They allow to consolidate at a single point the various social activities in such a way that the user is not required to log in each SNS and perform same social activity (Virmani, Pillai, & Juneja, 2014). The user performs a given social activity within a SNA and the information is synchronized to all of the social networks that the user specifies. Each sSNA provides the users with specialized features to integrate the SNSs but none of them tries to integrate the information available within SNSs. Compared to them, our proposed solution allows the users to not only collect data from different SNSs, but also to extract useful information.

2.2. Information filtering

Information filtering deals with the delivery of information that the user is likely to find interesting or useful. An information filtering system assists users by filtering the data source and delivers relevant information to them (Ghorab, Zhou, O'Connor, & Wade). Information filtering is very close with *Information Retrieval* and

[1] http://www.peekyou.com/.
[2] https://pipl.com/.
[3] https://developers.google.com/social-graph/.
[4] http://www.foaf-project.org/.
[5] http://www.sioc-project.org/.
[6] http://smiy.sourceforge.net/wi/spec/weightedinterests.html.
[7] http://openprovenance.org/.
[8] http://activitystrea.ms/.
[9] http://dbpedia.org/.
[10] http://www.opencalais.com/.
[11] http://friendfeed.com/.
[12] https://hootsuite.com/.
[13] http://sourceforge.net/projects/flock/.

Please cite this article in press as: Vu, X. T., et al. A user-centered and group-based approach for social data filtering and sharing. *Computers in Human Behavior* (2014), http://dx.doi.org/10.1016/j.chb.2014.11.079

inherits information retrieval research problems as well as results (Belkin & Croft, 1992). In the case of SNSs, information filtering focuses on the continuous analysis of the user's social streams and takes appropriate actions that either ignore or bring the information to the user's notice.

Different approaches devoted to social data have been proposed. One of them is to provide the user with contents about trending or hot topics. Such contents are determined based on how much they are shared and spread by the user's connections or by all members of the system (Cataldi, Di Caro, & Schifanella, 2010; Ishikawa et al., 2012). Another solution is to assist the users with the organization/categorization of their connections into significant groups or lists. This grouping task can be done using adapted interface (Zhang, Wang, & Vassileva, 2013) or automatically (Carmagnola et al., 2009; Qu & Liu, 2011; Rakesh, Singh, Vinzamuri, & Reddy, 2014). The users are then able to split and target the social data stream into more specialized sub-streams.

Social recommender systems have also been much studied as filtering solutions for social data. Compared to traditional recommender systems (Adomavicius & Tuzhilin, 2005), social recommender systems incorporate certain social data (e.g. trusted and untrusted users, followed and followers, friends lists, etc.) as contextual information to improve the personalized recommendation (Bobadilla, Ortega, Hernando, & Gutiérrez, 2013). Some SNSs have implemented their internal social recommender systems. For example, Facebook has used a specific ranking algorithm, named Edgerank, to personalize the user's new feeds. This algorithm scores every update based on three main metrics which are affinity (i.e. the interaction between two users), weight (i.e. the type of update) and time decay (i.e. the recency).[14] As a result, the updates of friends who interact the most with the user will go to the top of the stream while those of other contacts will be not shown much.

Some interesting Twitter-oriented academic words were also introduced in Sriram, Fuhry, Demir, Ferhatosmanoglu, and Demirbas (2010), Mendes, Passant, and Kapanipathi (2010), and Chen, Nairn, Nelson, Bernstein, and Chi (2010). Sriram et al. (2010) proposed to categorize tweets (i.e. short messages published by Twitter users) into some classes: (i) personal tweets/corporate tweets, (ii) event-related tweets, (iii) opinionated tweets, (iv) private tweets so that the user can choose what to see. The authors of Mendes et al. (2010) and Kapanipathi and Orlandi (2011) proposed to store all public tweets into a central repository, and annotate each tweet with one or several predefined topics using a dictionary. The users first need to select the topics of their interests and then regularly receive matching tweets. Chen et al. (2010) tried to build the user bag-of-words profile based on frequent words extracted from the user's tweets. Tweets containing URL posted by the user's friends were ranked against his/her profile before being suggested to the user.

The most important requirement of any information filtering systems, in particular recommender systems, is to efficiently build the user's interest/preference profile (Abdel-Hafez & Xu, 2013). For that purpose, the systems can explicitly ask the user for his/her interests or implicitly learn them from his/her activities. In the second case, the task is not obvious taking into consideration the facts that the interests of users in general do not follow a simple predictable model, and users have a wide range of interests across a large set of topics, even within a topic (May, Chaintreau, Korula, & Lattanzi, 2014). It is even less obvious when it comes to using large and often noisy social data for learning the user's interests. More advanced features such as temporal dynamics Abel, Gao, Houben, and Tao (2011) and Orlandi et al. (2012) must therefore be considered. Therefore, in our approach, the users are required to explicitly define their interests. Unlike in Mendes et al. (2010) and Kapanipathi and Orlandi (2011) where the topics were predefined, we let the users to freely add their own topics by means of specific queries. As we mainly focus on filtering textual data which arrive in large volumes, and are only relevant for very short time, we primarily applied some basic content-based information retrieval techniques such as boolean and vector space models for extracting contents.

2.3. Content curation

Content curation is the process of collecting, organizing and displaying information relevant to a particular topic or an area of interest. Content curation can be carried out either manually or automatically. More specifically, contents are first gathered from multiple sources using web-crawling or peer-sourcing and then associated with the corresponding topics, subtopics, and categories by designated people or automated programs.

There are many online content curation services, for example StumbleUpon,[15] Reddit,[16] Digg,[17] Pinterest.[18] Most of them allow the users to subscribe to predefined topics and provide the users with specific tools (i.e. browser plug-in) for annotating visited web pages with corresponding topics. The users somehow form a group linked to a given topic, where they can push and pull appropriate contents.

The group-based features of our proposed approach accordingly correspond to a semi-automatic content curation and sharing service. The interests of a group are collectively defined and updated by its members, while contents are automatically extracted from every member's social streams and categorized accordingly to a set of topics.

2.4. Discussion

There is a considerable number of solutions to Social Network Aggregation, information filtering, or content curation. However, most of them have so far been dedicated to one specific problem, and thus have not covered all of the three fields. Social Network Aggregators integrate multiple SNSs rather than integrate the information available within them. They facilitate users' social networking experience, but do not really ease their information filtering process. Retrieved data are simply put together without being filtered.

Information filtering solutions reduce users' information filtering efforts and provide interesting contents. However, most of them have applied different machine learning techniques which need to be trained by maximizing their performance on some set of training data. Consequently, they become domain-specific solutions. For example, an efficient solution dedicated to Twitter tweets may be not suitable for Facebook messages.

Content curation services allow users to collectively collect, and share interesting information with some group of interested people. Nevertheless, the members still have to manually select and push every piece of useful content to their different groups whereas their interesting contents published on SNSs are not considered.

The personalized social network aggregator and recommender named SocConnect Zhang et al. (2013) may be the most related to our work. It allows users to aggregate social data from Facebook and Twitter, and then help them to select useful contents. This

[14] http://whatisedgerank.com/.

[15] https://www.stumbleupon.com/.

[16] http://www.reddit.com/.

[17] http://digg.com/.

[18] http://www.pinterest.com/.

Please cite this article in press as: Vu, X. T., et al. A user-centered and group-based approach for social data filtering and sharing. *Computers in Human Behavior* (2014), http://dx.doi.org/10.1016/j.chb.2014.11.079

interesting work is however limited to individual basis and does not consider group perspective.

To our best knowledge, our work is the first attempt to re-adapt and combine the solutions from the three fields, Social Network Aggregation, information filtering, and content curation with the objective to help users to overcome the two identified problems: Information overload and "Walled-garden". Especially, we try to extend the user experience not only from one specific SNS to multiple SNSs, but also from an individual basis to a group perspective. Such extensions would allow users to take more advantage of their various social networks as well as their respective groups of interest. Furthermore, we apply a user-centered approach prioritizing the personalization features. This would make our proposed solutions more flexible and suitable to many users.

3. User-centered and group-based social data filtering & sharing

3.1. User-centered social data filtering

In this subsection, we introduce a user-centered solution for tackling the two previously cited problems: information overload and "walled garden". We first present a common model for representing and integrating social data from different SNSs, then describe in details the process of information filtering.

3.1.1. Social Data Aggregation

The types of user information/properties covered by SNSs are different from one another. For example, the user profile on Twitter is currently very bare. It only includes name, bio and location of the member. In contrast, the user profile on Facebook is much elaborated. It includes: basic information such as the name, photo, age, birthday, relationship status, etc.; personal information such as interest, favorite music & TV shows, movies, books; and education and work information such as the names of schools attending/attended, and current employer. Furthermore, each SNS utilizes its own syntax and terms for representing user data, and often different terms for same type of data. For example, a piece of text published by a user is called "tweet" on Twitter but "post" on Facebook.

Given such diversity, a common model for representing social data should be generic as well as extensible to support frequent social data available in SNSs and to easily be extended to accept new kinds of social data. We have closely studied the top SNSs, namely Facebook, Twitter, LinkedIn, Google+, for available social data. From this study, we determined the six most frequent dimensions as follows:

1. The *Profile Information* dimension includes basic information about the user such as name, description, city, email, gender, location, etc.;
2. The *Friends* dimension represents connections established between the user and others;
3. The *Groups* dimension contains information about the groups in which the user is involved;
4. The *Studies* & *Works* dimension describes respectively the school and academic experience and the professional experience of the user;
5. The *Interests* dimension lists the user's interests;
6. The *User-created Contents* represents all contents created by the user.

It is important to note that these dimensions are not completely exclusive to each other and there may be some overlap between them. For example, people often join a specialized group because they share with the group some common interests, but also a dedicated group formed by people from the same school or the same workplace.

Based on these dimensions, we have therefore built our generic model. As shown in Fig. 1, a user can have several *social accounts* which contain the properties identical to the *Profile Information* and the *Studies* & *Works* information. Each social account is associated with a number of timestamped *social data* through different *social activities* which help to categorize these associations. There are at this time three types of social data, *member*, *post*, *interest*. They are linked to a social account via four different social activities: (i) a social account posts a post, (ii) it receives a post posted by another member, (iii) it befriends with a member, (iv) it adds a new interest. We combined the two dimensions *Interests* and *Groups* under the interest type of social data. We also extracted *links* out of posts as several posts may refer to a same link.

As we mainly focus on textual data and apply content-based retrieval technique for filtering social data, every subclass of social data should have at least one text-valued property. For example, a member shall include a description. An interest shall include a name and a description. A post shall contain a text and eventually the title and description of the linked url.

This generic model is simple to extend. For example, if we identify a new and mandatory type of social data, we can then add it as a subclass to the social data class and define the corresponding social activity. It is also easy to select the different views of a user's aggregated social data which are the profile information, the friend list, the posts (published by him/herself), the following posts (published by his/her friends), and the list of interests, for processing and displaying.

3.1.2. Information filtering

The information filtering process consists of constantly analysing any new social data and accordingly taking appropriate actions that either ignore or show it to the user. Thus, it should know about the user's information needs which reflect the user's short, medium or long-term interests (Ghorab et al.). Such information can be gathered in an implicit manner where it is obtained without any extra effort from the user or in an explicit manner where the users have to explicitly supply information to the system (Gauch, Speretta, Chandramouli, & Micarelli, 2007).

Unlike domain-specific recommender systems where data are quite homogeneous, social data are much more diverse. Java, Song, Finin, and Tseng (2007) have shown four types of social data: (i) daily chatting about daily routine or what people are currently doing, (ii) conversations to comment or reply to their friends' posts, (iii) information/URLs sharing, (iv) news reporting to report latest news or comment about current events. Moreover, people are often influenced by their social connections, so change their interests regularly. All of this makes it challenging to efficiently learn a user's interests in an implicit manner from his/her social data.

Therefore, in our approach, the users explicitly specify their information needs which reflect rather their medium or long-term interests than short-term interests. In other words, the information needs are not spontaneous like when searching with search engines but last for certain time. The users are free to add the topics of their interest which could be very general or specialized, and define them using specific queries called *selector*. There are three types of selector:

1. *Hashtag-based selector* expects as value a valid *hashtag* which is a word or a phrase prefixed with the symbol "#". As hashtags are already widely used in SNSs, particularly on Twitter and Facebook, to collectively group and efficiently retrieve messages (Murtagh, 2013), using hashtag-based selector will be very natural to users. The hashtag-based selection is in fact an

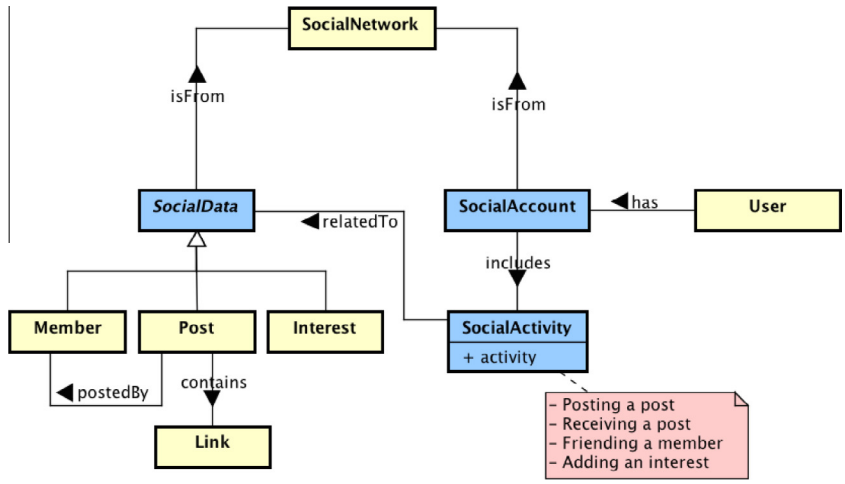

Fig. 1. The proposed common model for integrating social data across SNSs.

"upstream" effort. When posting some content in SNSs, the user can insert with it the chosen hashtag to indicate its relevancy. So, content containing the chosen hashtag will be directly selected as relevant.

2. *Keyword-based selector* follows the same principle of web search query and thus allows the users to create queries with boolean operators (i.e. "AND", "OR", "NOT"). It will moreover expand the initial query with its derived forms and/or its synonyms using dedicated dictionaries.
3. *Concept-based selector* requires a semantic concept belonging to an ontology publicly accessible (e.g. DBpedia – the semantic database of Wikipedia). Compared to keyword-based selectors, it is more powerful for matching information. Firstly, it provides multi-languages labels for a single concept. Secondly, it allows to expand the given concept by its related concepts, thus to disambiguate the concept and to improve the matching precision as well.

While, the first and the second types of selector are already very common to the users, the concept-based selector is more complex and less intuitive. With a given topic, a user can associate as many selectors as he/she desires in order to increase the probability for detecting interesting information. The social data will be automatically matched against the user's selectors. Matching information will then be organized under the corresponding topics. This way, the user can easily browse and view interesting information by topics. Note that a given piece of information may correspond to several topics.

In brief, our user-centered social data filtering provides the users with a centralized access to interesting information available within SNSs. By creating a number of topics, the users can split and target their social streams into several topic-related sub-streams. Our semi-automatic filtering solution is generic enough to cover the users' heterogeneous social data. Furthermore, the user-friendly methods like hashtags or keywords do not require users to have specific knowledge.

3.2. Group-based social data sharing

Although the previously proposed solution can help the user to overcome the two identified problems, it limits the user experience only to his/her personal streams. On the one hand, there may be other useful social data outside of the user's circle of friends which remain unnoticed to the user. On the other hand, people are often members of different communities held by common interests. Some join a group because they felt the urge to contribute to the cause; others come because they can benefit from being part of the group. There is however no guarantee that a user's circle of friends overlaps the members of his/her different groups. Thus, interesting information published by a member in his/her subscribed SNSs, is not always visible to the other members of the groups. It would be useful and profitable for the members of a group to be able to share interesting information with no more effort required than it takes to share information in SNSs. Based on this assumption, we therefore extended the proposed solution for supporting group-based social data sharing.

3.2.1. Private and open groups

Until now, matching information is organized under topics that the users can directly access and view. This is sufficient for personal information management but not enough for sharing. Thus, we introduced an extra level of organization that we called *group*. There are two categories of group: (i) private groups and (ii) open groups. The former is only accessible by its creator whereas the latter is an open space for people to join, and to share their social data with others. One user can be a member of several private and open groups if he/she want to.

In the case of open groups, the users may not want to share all of their social data as some could be sensible or undesired information. We therefore provide the users with features so that they can customize their sharing experience. The users are encouraged to adapt for each of their groups the three following settings (see the *memberOf* relation in Fig. 2):

- *Authorized accounts* indicates which social accounts and their associated social data can be matched and eventually shared with the group;
- *Authorized data* indicates which views of social data can be matched and eventually shared with the group (e.g. the posts and the following posts but not the friend list);
- *Review* if enabled, prevents a relevant social data to be immediately shared with the group but waiting for the user's approval.

3.2.2. Personalized sharing

After becoming a member of a given group, the user can suggest a new topic or a new selector associated with a defined topic whenever he/she finds it relevant to the group. The user is also free to choose to follow the topics and even to only accept a subset of their associated selectors which seem to be the most relevant to his/her own interests. Such a personalization feature is very important,

Please cite this article in press as: Vu, X. T., et al. A user-centered and group-based approach for social data filtering and sharing. *Computers in Human Behavior* (2014), http://dx.doi.org/10.1016/j.chb.2014.11.079

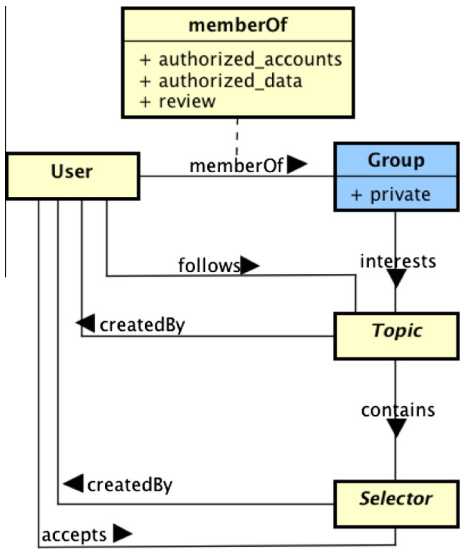

Fig. 2. Sharing experience personalization.

because it prevents the members from accepting systematically all topics and receiving all matching information. Otherwise, the users will be facing again the problem of information overload.

The user's sharing settings and personalized topics will be taken into account during the filtering process. In other words, the filtering process is also personalized. Only authorized data will be selected from the user's social data for matching against only the selectors that the user has accepted. Non-matching information will never be shared with the group.

3.2.3. Group's collective intelligence

The main benefit of this group-based social data sharing is to promote the group's collective intelligence. In other words, it encourages every member to contribute his/her own intelligence to improve the information sharing within the group. First of all, it allows to grow the number of information sources. These sources are more or less "reliable" based on two assumptions: (i) a person who is interested in a topic often tries to share interesting information or useful content in order to influence his/her friends, (ii) people with similar interests are more likely to be connected. Such assumptions can be explained by social correlation theories such as homopholy (McPherson, Smith-Lovin, & Cook, 2001) and social influence (Marsden & Friedkin, 1993).

If this is an implicit way of contribution, the members of a group can also explicitly contribute to define the group's topics and selectors. Three types of contributions are possible: (i) adding a new topic as early as it becomes a trending topic in SNSs, (ii) adding more precise, advanced selectors to improve the filtering precision, and (iii) rejecting (by not accepting) inadequate topics or selectors so that they should disappear from the group. If the members are active enough, the group will have enriched, updated topics and precise selectors, and thus end with more appropriate contents.

Another important collective effort is to agree on using some representative hashtag-based selectors. Thereby, each member will become an active and efficient selector for capturing the best information and the most useful contents for the group.

4. Extensible system architecture

In this section, we present a system architecture for implementing the proposed approach. The whole system is broken down into important modules, as illustrated in Fig. 3, so that it is simple to implement as well as to improve. We detail below the role of each module, its current implementation, and its eventual issues and improvement possibilities respectively.

4.1. Social Data Aggregation

The *Social Data Aggregation* module is composed of a number of dedicated aggregators responsible for aggregating the user's social data from different SNSs. The first and mandatory step is to ask the user for authenticating and authorizing the aggregators according to the authentication and authorization protocols of the corresponding SNSs. Once granted, the aggregators will be able to request different APIs (e.g. Facebook Graph API,[19] Twitter Rest API[20]) for collecting the users' social data. Each aggregator contains hard-coded mapping rules of the data model handled by the SNS against our data model, based on which the aggregator can request for the needed social data and convert the received data into some expected syntaxes. We have re-adapted the open source library HybridAuth[21] which acts as an abstract API between our system and various proprietary APIs. Hence, we only needed to extend existing classes by specifying the types of needed social data and including mapping rules. We have moreover programmed the aggregators in such a way that they can auto-run at regular time intervals (e.g. several times per day) to retrieve the user's recent social data.

The biggest issue of this solution lies in the fact that it relies on no single standard, but multiple formats provided by different SNSs. The aggregators should be reviewed to respond to any change in format whenever it arises. Another possible drawback is that social data are only gathered at regular time intervals. Recent interesting information may still be ignored by the system as the users log in. To cope with this problem, we could use the real-time update features provided by certain SNSs to receive new data within a couple of minutes of their occurrence.

4.2. Social Data Storing

The *Social Data Storing* module is responsible for storing social data aggregated by the *Social Data Aggregation* module. It constantly inserts new data and frequently receives a lot of requests from other modules. We have implemented this module with a relational database (i.e. MySQL).

This type of database can be a critical factor for the system performance as the number of users and subsequently their social data increase. To improve the performance, we could migrate data to other storing system models like NoSQL databases which are designed for scalability.

4.3. Social Data Enrichment

The *Social Data Enrichment* module attempts to enrich social data in order to improve the effectiveness of the filtering process. However, this is not a trivial task as social data often contain very little text and are ungrammatically written. We have applied one technique that consists of appending social data containing external links with the content extracted from the referred web pages (e.g. the titles and the descriptions). This simple technique seems to be very useful, since lots of social data contain Web links.

More sophisticated methods for text processing, for example *entity extraction* or *text classification*, have been tested with social data in Gao, Abel, and Houben (2012) and Orlandi et al. (2012). Some of them could be applied to better enhance the enrichment.

[19] https://developers.facebook.com/docs/graph-api.
[20] https://dev.twitter.com/docs/api/1.1.
[21] http://hybridauth.sourceforge.net/.

Please cite this article in press as: Vu, X. T., et al. A user-centered and group-based approach for social data filtering and sharing. *Computers in Human Behavior* (2014), http://dx.doi.org/10.1016/j.chb.2014.11.079

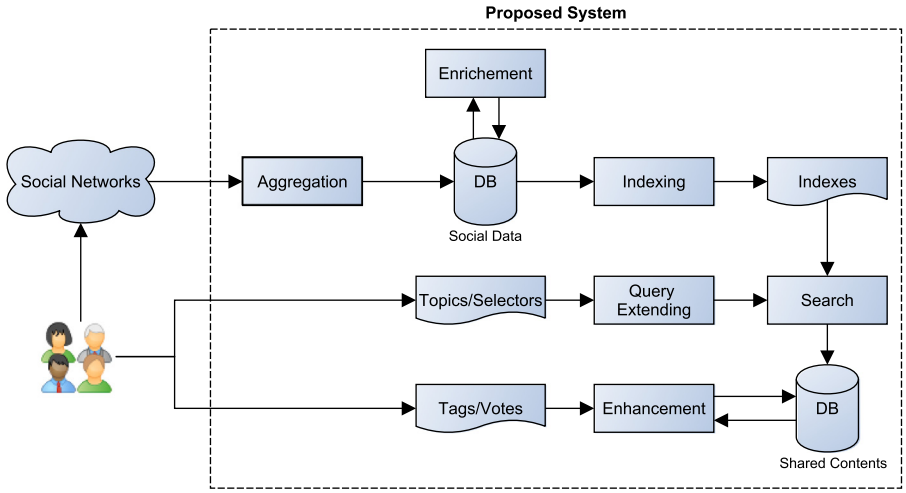

Fig. 3. System architecture for implementing the proposed approach.

4.4. Social Data Indexing

The role of *Social Data Indexing* module is to index aggregated social data in an incremental manner. It extracts and saves the index terms from a piece of social data as soon as it is gathered. The extracted index terms later serve as input for filtering information. We have implemented this module by using Lucene[22] which is a very well-known open source information retrieval library.

One of the biggest difficulties is to cope with multi-language textual data. General-purpose text analysers like Lucene's standard analyser that we have used, though relevant to English and certain non-English languages (e.g. French), may give poor results to other languages. To improve this, we could add an extra step that detects the language in which the texts are written. Given the language, the system will choose and use a language-specific analyser.

4.5. Query Expansion

The *Query Expansion* module translates a selector into the internal queries according to its type (see the example in Fig. 4). For hashtag-based selectors, there is nothing to be done, since the same values will be used. For keyword-based selectors, we have extended the initial value by its derived forms (e.g. singular, plural) and/or its synonyms based on dedicated dictionaries. For concept-based selectors, we have retrieved the multi-language labels of the given concept by requesting the corresponding dataset.

Concept-based selectors would be further expanded if we could associate the given concept with its closely related concepts in order to disambiguate it. For example, "city" may be appended to "Paris" (i.e. "Paris" AND "city") to clearly indicate that the results should be in relation to the capital of France.

4.6. Information filtering

The *Information filtering* module constantly analyses new social data and organizes them under different topics within various groups. With respect to the personalization principle, this module has to follow the following steps:

1. For each group, get its member list.
2. For each member.
   (a) Get his/her sharing settings (i.e. authorized accounts, authorized data).

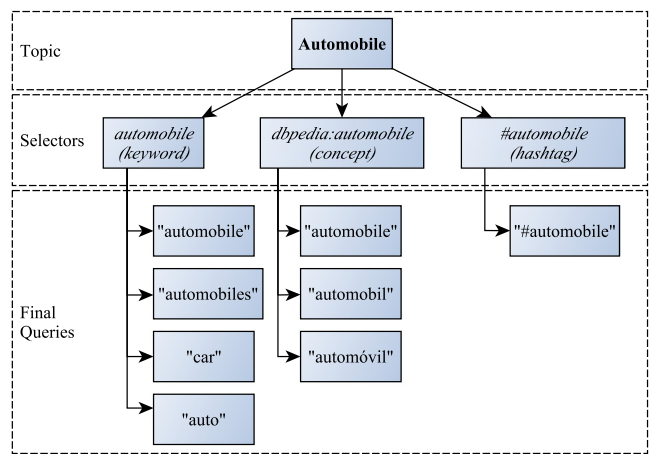

Fig. 4. Example of Query Expansion, the final queries of a selector will be appended into a unified query by the "OR" operator.

   (b) Get the selectors that the member has accepted.
   (c) Search the accepted selectors against the index terms of new authorized social data.
   (d) For each matching social data, associate it with the corresponding topics.
3. Save all matching social data as the group's shared contents.

To compute the topic relevancy of social data, we have combined two well-known methods, the *Boolean model* and the *Vector Space Model*, of information retrieval (Manning, Raghavan, & Schütze, 2008). Each piece of social data, considered as a document, has been first matched against a boolean model and then scored by a vector space model by the boolean model and then scored by the vector space model. Other more sophisticated methods such as those being reviewed in Ghorab et al., could be studied to improve the information filtering process.

4.7. Enhancement

The *Enhancement* module tries to improve the group's shared contents based on the collaborative efforts (i.e. the members' explicit feedbacks). The members can vote a piece of content as relevant or irrelevant or tag it to additional topics or selectors. The point is that the entire group can benefit from the work of some members.

[22] http://lucene.apache.org/core/.

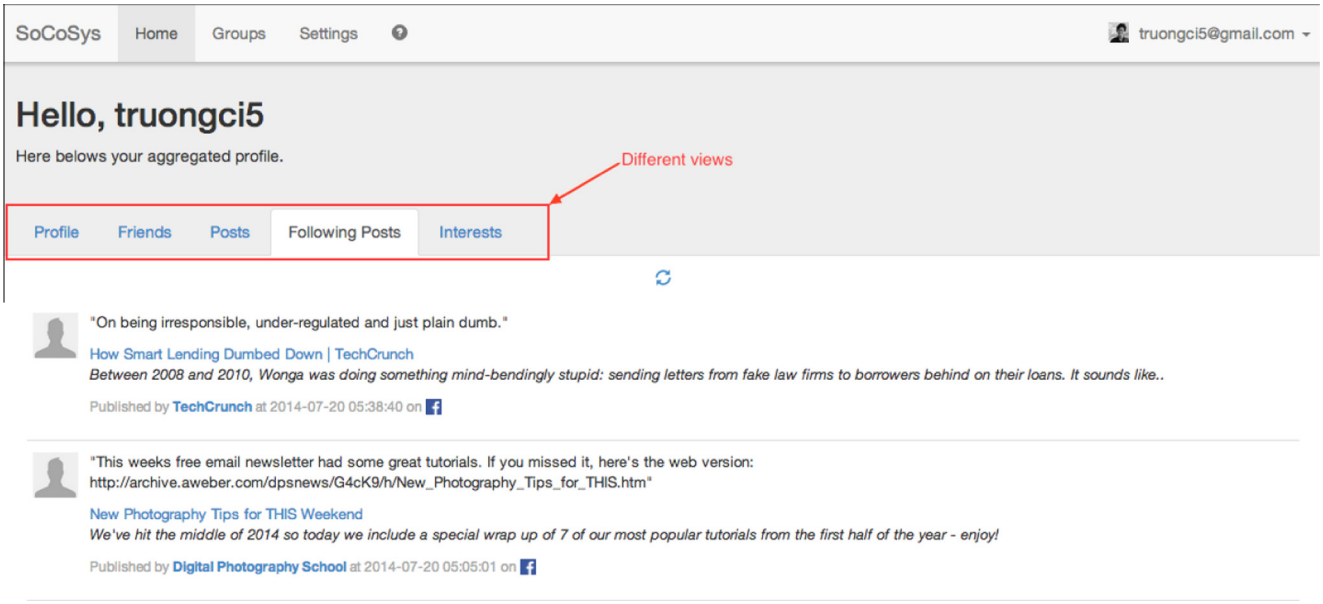

Fig. 5. The user's aggregated data.

If a piece of content receives a certain number of irrelevant votes, it will be removed off the group. If there are some members tagging a piece of content with a same additional topic or selector, the topic or selector will be definitively associated to the content.

In both cases, the suitable thresholds have to be defined. For now, we have roughly set them at some static values, as we only have small groups. Dynamical techniques will be needed to cover groups with various sizes.

4.8. Advanced features

Some other advanced features can also be added into the system. One of them is *group analysis* which allows to identify the topic-based experts within a group and to go deep into the evolution of the group's areas of interest over time. Such features strengthened by visualization means could be very interesting for group-oriented decision making.

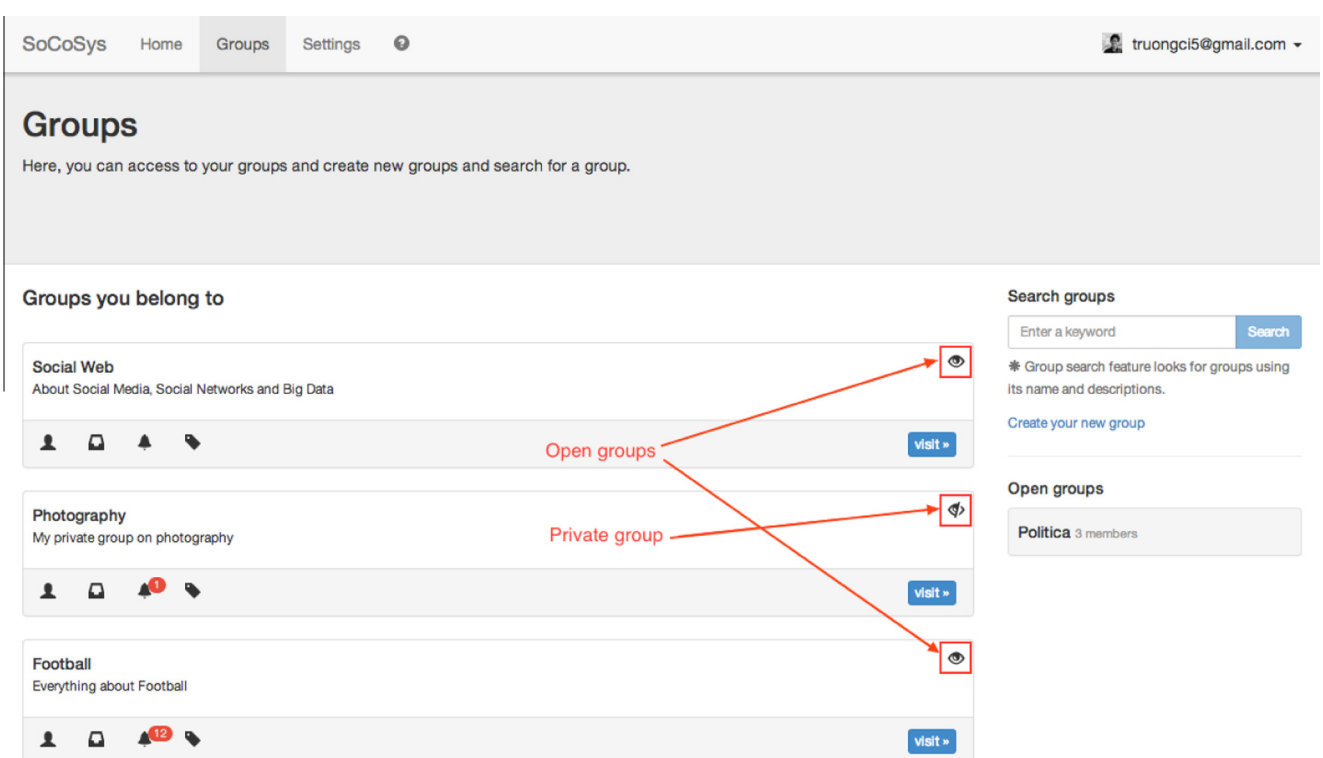

Fig. 6. The user's groups.

Please cite this article in press as: Vu, X. T., et al. A user-centered and group-based approach for social data filtering and sharing. *Computers in Human Behavior* (2014), http://dx.doi.org/10.1016/j.chb.2014.11.079

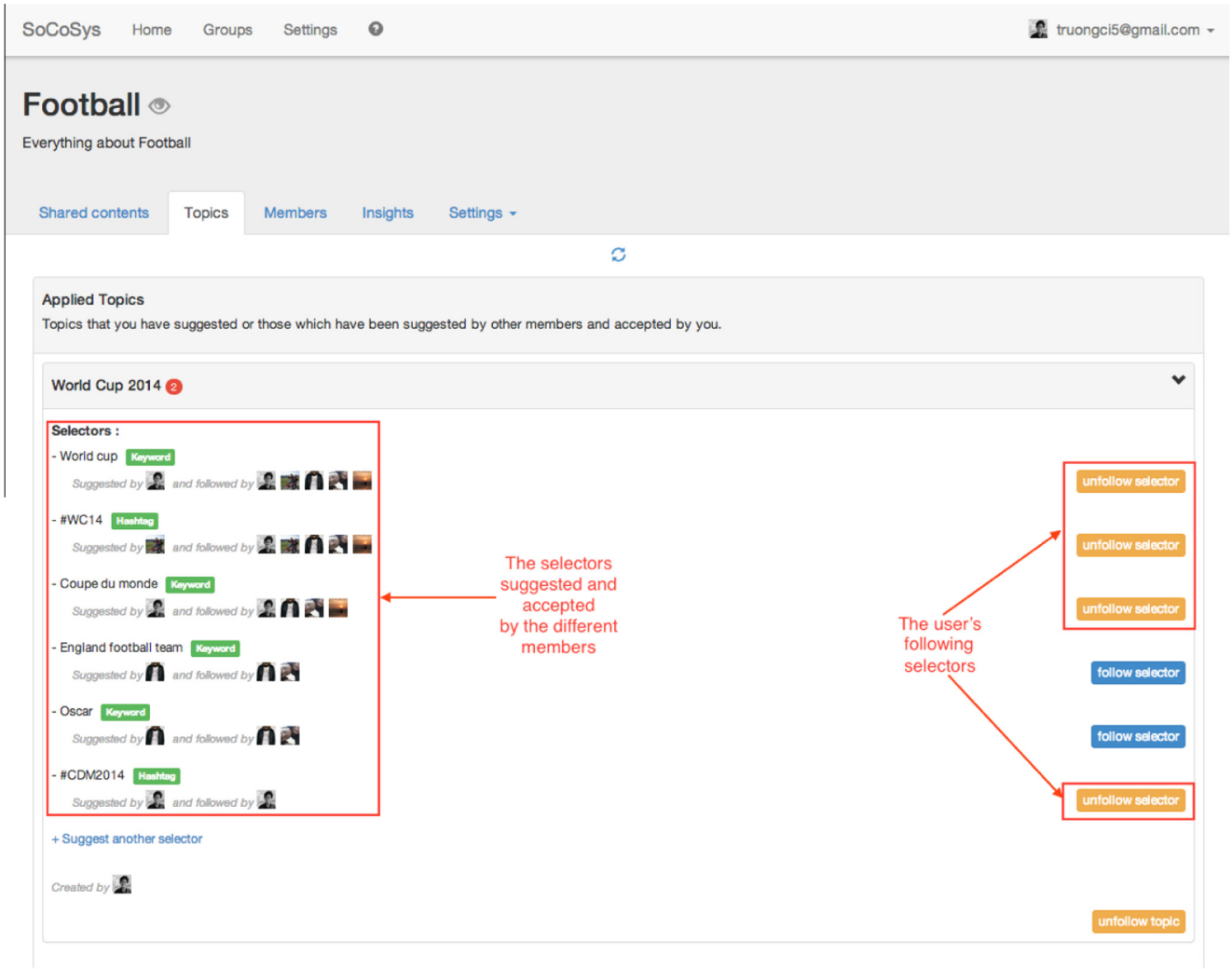

Fig. 7. The group's topics and selectors.

5. Web-based prototype and primary test

5.1. Web-based prototype

Based on the previous system architecture, we have implemented a first prototype called *SoCoSys* standing for *social and collective system*. This web-based service allows the user to register with a unique email, then to create an aggregated profile. The user can connect his/her Facebook, LinkedIn, Twitter profiles to build his/her aggregated profile. Once this has been done, the user can see his/her aggregated and updated social data arranged into five views: profile information, friend list, posts, following posts and interest list (see Fig. 5). Such aggregated profile is exclusive to the user, no one else can see it.

Then, the user can create, search and join a group (see Fig. 6). The user can choose to create open or private groups. However, it is only possible to search and join open groups, since private groups are exclusive to their creators. Special notifications like the number of new members, the number of new topics or the number of last detected contents, give a good indication for the user to decide which groups to visit first.

Once the user joined a group, he/she is encouraged to adapt the sharing settings and to customize his/her interests. The user may decide whether or not to follow the topics suggested by other members, also choose for each followed topic, appropriate selectors (see Fig. 7). Of course, the user can propose his/her own topics and selectors as well.

Finally, the user can view all matching contents in a chronological order or further filter contents by selecting a particular topic (see Fig. 8).

5.2. Primary test

We invited a small group of users to test this prototype. The test is not meant to represent a real experiment, but to observe how the users understand the proposed approach and utilize the provided services. The test group consisted of ten volunteered international PhD students at the University of Technology of Compiègne, and regular users of SNSs. They were introduced to our social data filtering approach with its two modes of use (i.e. personal and group-based) and to the operation and specific features of the prototype.

After one month of observation, we found out that:

- All participants granted access to their Facebook profiles, but only 6 of them connected at least one another profile.
- On average, they had more than 300 friends on Facebook and received via these connections nearly 180 contents per day.

Please cite this article in press as: Vu, X. T., et al. A user-centered and group-based approach for social data filtering and sharing. *Computers in Human Behavior* (2014), http://dx.doi.org/10.1016/j.chb.2014.11.079

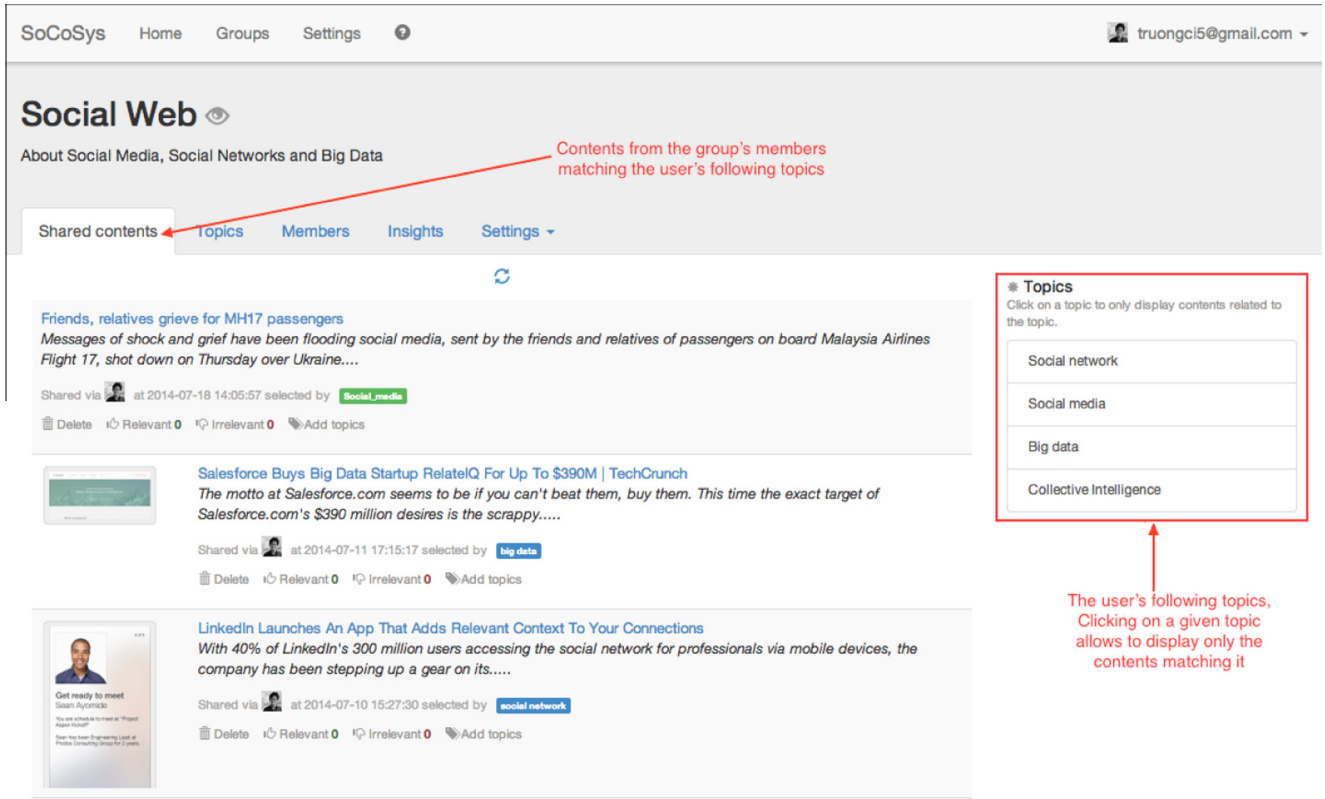

Fig. 8. The group's shared contents.

- 8 out of 10 participants joined open groups and most of them accepted to open all their social data with their groups and also disabled the *review* option. This could be explained by the fact that they already knew and trust each other.
- There were in total 6 private groups and 4 open groups.
- The centers of interest of the open groups varied a lot from general areas like *Football*, *Politics* to specialized areas like *Social Media* and *Social Responsibility*.
- The two most "successful" groups were *Football* and *Politic* which had respectively 6 and 4 members.
- Within the open groups, the participants mainly created topics following some major real events, for example, "the FIFA world cup" or "the Brazilian general election" while in the private groups, they preferred more static topics, for example "Photography" or "Guitar".
- The members of the open groups were quite active for suggesting selectors. They mostly used keyword-based selectors and hashtag-based selectors, with which they are already familiar. Moreover, they did not accept systematically every selector but selected well those corresponding to their interests.
- There are two behaviours when suggesting selectors: (i) adding multi-language or synonymous terms, (ii) adding specialized terms. For example, in the case of "the FIFA world cup", while some added three keywords "world cup", "coupe du monde" and "copa do mundo" to be able to follow the event in different languages, others added "England football team" to keep track of the specific element of the event.
- Finally, although we have not properly evaluated the effectiveness of the information filtering process yet, the participants gave generally positive feedbacks about the number and the quality of matching contents, especially in the case of private groups.

These interesting findings have confirmed our initial assumptions for a user-centered and group-based approach for social data filtering and sharing. The users have been clearly interested by a solution for helping them to extract information relevant to their various interests from their numerous social data. They have also found the interest of collaboration (i.e. sharing, learning) with others in a group setting.

Moreover, this test gave rise to some additional functionalities that we should study and eventually include in the next version of SoCoSys, for example:

- *Duplicated information hiding*: Several matching contents, especially in open groups, can refer to the same information, even though they have been shared by different people and/or published by different sources. On the one hand, this means that the information is important. On the other hand, the users will probably be bored by viewing similar contents. Thus, we should investigate means to highlight the important information while hiding the duplicated information
- *Sharing with SNSs*: The users discover new and interesting contents within their respective groups. In some cases, one may find a given piece of content particularly interesting and feel the need to share it with, for example, his Facebook friends. It could be done easily by using the APIs provided by SNSs. However, we should add additional filters to prevent these contents from being selected once again by SoCoSys.
- *Notification*: The users are busy, and thus can not regularly visit SoCoSys. To help the users to stay current with interesting information, we could include an email notification feature. It, with a personalized frequency, notifies the users of the new activities (e.g. the newly detected contents, the newly added topics, etc.), if any.

6. Conclusion

We have shown, in this paper, a user-centered and group-based approach for social data filtering and sharing. This approach contains

a user-centered solution for addressing two problems of today SNSs: information overload and "walled garden". This solution allows the users to aggregate social data across SNSs and filter, organize them under topics. For aggregating diverse social data, we introduced a generic model which is able to integrate the most frequent information dimensions of social data available in different SNSs and more importantly, is easy to be extended. We also applied a semi-automatic information filtering process. The users first explicitly define their topics of interest using three types of specific query: (i) keyword-based selector, (ii) hashtag-based selector and (iii) concept-based selector. Social data are then filtered and matching information will be organized under topics.

The proposed approach also presents a new group-based use of social data to go beyond its primary personal purpose. This special feature allows the users to share some part of their aggregated social data with their respective groups, thus access to more interesting information and useful contents. Specially, the users can personalize their sharing experience by editing their sharing settings and participating in the definition of the group's interests. We furthermore discussed the major benefits of such collaboration in terms of collective intelligence promotion.

Beside these two main contributions, an extensible system architecture for implementing our approach has also been presented. This architecture is made up of a number of important modules which are responsible for significant steps within the whole process: social data aggregating, social data storing, social data enrichment, social data indexing, query expansion, information filtering, enhancement. For each module, we have also discussed important issues and possible improvements.

We have implemented a first web-based prototype called SoCoSys supporting the three SNSs, namely Facebook, Twitter, and LinkedIn. A primary test using SoCoSys has thus been carried out with a small group of volunteered participants. The findings have shown encouraging results and motivated us to work on a final prototype. A group of computer engineering students from University of Technology of Compiegne,[23] and a group of employees from the 50A company [24] will be asked for testing this version of SoCoSys. The first group will use the system as a technology watch tool while the second group will use it for content curation. The output will allow us to better assess the proposed approach.

It is finally worth mentioning that our proposed approach composed of generic components is totally generalizable. It is not limited to any specific SNS or to any specific application domain. Our current implementation (SoCoSys) is also general-purpose. On the other hand, it is possible to re-adapt, extend the proposed modules within another implementation to better contextualize our approach.

Acknowledgement

We would like to thank the 50A company (www.50a.fr) for funding this work.

References

Abdel-Hafez, A., & Xu, Y. (2013). A survey of user modelling in social media websites. *Computer and Information Science, 6*(4), 59–71. http://dx.doi.org/10.5539/cis.v6n4p59.

Abel, F., Gao, Q., Houben, G., Tao, K. (2011). Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web, In *Proceedings of the ACM WebSci'11, Koblenz, Germany* (pp. 1–8).

Adomavicius, G., & Tuzhilin, a. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734–749. http://dx.doi.org/10.1109/TKDE.2005.99.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM, 35*(12), 29–38. http://dx.doi.org/10.1145/138859.138861.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems, 46*(0), 109–132. http://dx.doi.org/10.1016/j.knosys.2013.03.012.

Bonds-Raacke, J., & Raacke, J. (2010). MySpace and Facebook: Identifying dimensions of uses and gratifications for friend networking sites. *Individual Differences Research, 8*(1), 27–33.

Borgs, C., Chayes, J., Karrer, B., & Meeder, B. (2010). Game-theoretic models of information overload in social networks. In R. Kumar & D. Sivakumar (Eds.), *Algorithms and models for the web-graph* (pp. 146–161). Berlin Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-18009-5_14.

Buccafurri, F., Lax, G., Nocera, A., Ursino, D. Moving from social networks to social internetworking scenarios: The crawling perspective. Information Sciences. doi: http://dx.doi.org/10.1016/j.ins.2013.08.046.

Carmagnola, F., Vernero, F., & Grillo, P. (2009). Sonars: A social networks-based algorithm for social recommender systems. In H. Geert-Jan, M. Gord, P. Fabio, & M. Zancanaro (Eds.), *User modeling, adaptation, and personalization* (pp. 223–234). Berlin Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-02247-0_22.

Cataldi, M., Di Caro, L., & Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining, MDMKDD '10* (pp. 4:1–4:10). New York, NY, USA: ACM. http://dx.doi.org/10.1145/1814245.1814249.

Cavazza, F. (2014). Social media landscape 2014. <http://www.fredcavazza.net/2014/05/22/social-media-landscape-2014/>.

Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. (2010). Short and tweet: Experiments on recommending content from information streams. In *Proceedings of the SIGCHI conference on human factors in computing systems, CHI '10* (pp. 1185–1194). New York, NY, USA: ACM. http://dx.doi.org/10.1145/1753326.1753503.

Chisari, M. (2009). The future of social networking. In *W3C workshop on the future of social networking*, Barcelona.

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM, 57*(6), 74–81. http://dx.doi.org/10.1145/2602574.

Fire, M., Puzis, R., & Elovici, Y. (2013). Organization mining using online social networks. *CoRR, abs/1303.3*, 1–19. arXiv:arXiv:1303.3741v1.

Gao, Q., Abel, F., & Houben, G.-J. (2012). GeniUS: Generic user modeling library for the social semantic web. In *Proceedings of the 2011 joint international conference on the semantic web* (pp. 160–175). Berlin, Heidelberg: Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-29923-0_11.

Gao, Y., Wang, F., Luan, H., Chua, T.-S. (2014). Brand Data gathering from live social media streams. In *Proceedings of international conference on multimedia retrieval – ICMR '14* (pp. 169–176). doi: http://dx.doi.org/10.1145/2578726.2578748.

Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). User profiles for personalized information access. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web* (pp. 54–89). Berlin, Heidelberg: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-72079-9_2.

Ghorab, M.R., Zhou, D., O'Connor, A., Wade, V. Personalised information retrieval: Survey and classification. User Modeling and User-Adapted Interaction. doi: http://dx.doi.org/10.1007/s11257-012-9124-1.

Ishikawa, S., Arakawa, Y., Tagashira, S., & Fukuda, A. (2012). Hot topic detection in local areas using Twitter and Wikipedia. In *Proceedings of the workshops of the international conference on architecture of computing systems*. München, Deutschland: IEEE Computer Society.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we Twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis* (pp. 56–65). San Jose, CL: Springer.

Kapanipathi, P., Orlandi, F. (2011). Personalized filtering of the twitter stream. In *SPIM workshop at* ... (pp. 1–8).

Kim, W., Jeong, O.-R., & Lee, S.-W. (2010). On social web sites. *Information Systems, 35*(2), 215–236. http://dx.doi.org/10.1016/j.is.2009.08.003.

Kwak, H., & Lee, H. G. (2011). A review of research on social network services using the new media evolutionary model. *Informatization Policy, 18*(3), 3–24.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York, NY, USA: Cambridge University Press.

Marsden, P. V., & Friedkin, N. E. (1993). Network studies of social influence. *Sociological Methods & Research, 22*(1), 127–151. http://dx.doi.org/10.1177/0049124193022001006.

May, A., Chaintreau, A., Korula, N., & Lattanzi, S. (2014). Filter & follow: How social media foster content curation. *SIGMETRICS Performance Evaluation Review, 42*(1), 43–55. http://dx.doi.org/10.1145/2637364.2592010.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415–444. http://dx.doi.org/10.1146/annurev.soc.27.1.415.

Mendes, P. P. N., Passant, A., & Kapanipathi, P. (2010). Twarql: Tapping into the wisdom of the crowd. In *Proceedings of the 6th international conference on semantic systems, I-SEMANTICS '10* (pp. 45:1–45:3). New York, NY, USA: ACM. http://dx.doi.org/10.1145/1839707.1839762.

Meo, P. D., Nocera, A., Terracina, G., Ursino, D., & De Meo, P. (2011). Recommendation of similar users, resources and social networks in a Social Internetworking Scenario. *Information Sciences, 181*(7), 1285–1305. http://dx.doi.org/10.1016/j.ins.2010.12.001.

[23] http://www.utc.fr/.

[24] http://www.50a.fr/.

Murtagh, R. (2013). The role of #hashtags in social media and search. <http://searchenginewatch.com/article/2305444/The-Role-of-Hashtags-in-Social-Media-and-Search>.

Orlandi, F., Breslin, J., & Passant, A. (2012). Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th international conference on semantic systems* (pp. 41––48). New York, NY, USA: ACM. http://dx.doi.org/10.1145/2362499.2362506.

Qu, Z., & Liu, Y. (2011). Interactive group suggesting for Twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: Short papers, HLT '11* (pp. 519–523). Stroudsburg, PA, USA: Association for Computational Linguistics. Vol. 2.

Rakesh, V., Singh, D., Vinzamuri, B., & Reddy, C. C. K. (2014). Personalized recommendation of Twitter lists using content and network information. In *Proceedings of the eighth international AAAI conference on weblogs and social media* (pp. 416–425). Michigan, USA: The AAAI Press.

Smith, A. (2014). 6 new facts about facebook. <http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/>.

Smith, C. (2014). 6 new facts about facebook. <http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/>.

Solis, B. (2013). The conversation prism by brian solis and jess3. <https://conversationprism.com/>.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval – SIGIR '10* (pp. 841–842). New York, USA: ACM. http://dx.doi.org/10.1145/1835449.1835643.

Virmani, C., Pillai, A., Juneja, D. (2014). Study and analysis of social network aggregator. In *Optimization, reliabilty, and information technology (ICROIT), India* (pp. 145–148). doi: http://dx.doi.org/10.1109/ICROIT.2014.6798314.

Zhang, J., Wang, Y., & Vassileva, J. (2013). SocConnect: A personalized social network aggregator and recommender. *Information Processing and Management, 49*, 721–737. http://dx.doi.org/10.1016/j.ipm.2012.07.006 (2012).