Par **Xuan Truong VU**

User-centered and group-based approach for social data filtering and sharing

Thèse présentée
pour l'obtention du grade
de Docteur de l'UTC

Université de Technologie de Compiègne

UMR CNRS 7253 Heudiasyc

Thèse

en vue de l'obtention du grade de

Docteur de l'Université de Technologie de Compiègne

Champ Disciplinaire : Technologie de l'Information et des Systèmes

User-centered and Group-based Approach for Social Data Filtering and Sharing

par

Xuan Truong Vu

Thèse soutenue le 1 avril 2015 devant le jury composé de :

Mme	Kathia Oliviera	Rapporteuse
M.	Pascal Molli	Rapporteur
Mme	Catherine Faron Zucker	Examinatrice
M.	Jean-Paul Barthes	Examinateur
Mme	Marie-Hélène Abel	Directrice de thèse
M.	Pierre Morizet-Mahoudeaux	Directeur de thèse
M.	Nicolas Bermond	Membre invité

"Continuous effort - not strength or intelligence - is the key to unlocking our potential."

Winston Churchill

"The secret of getting ahead is getting started. The secret of getting started is breaking your complex overwhelming tasks into small manageable tasks, and then starting on the first one."

Mark Twain

Abstract

User-centered and Group-based Approach for Social Data Filtering and Sharing

by Xuan Truong Vu

The social media have played an increasingly important role in many areas of our every day life. Among others, social network sites such as Facebook, LinkedIn, Twitter and Google+ have recently exploded in popularity by attracting millions of users, who communicate with each other, share and publish information and contents at an unprecedented rate. Besides the recognized advantages, social network sites have also raised various issues and challenges. We are particularly interested in two of them, information overload and "walled gardens". These two problems prevent the users from fully and efficiently exploiting the wealth of information available on social network sites. The users have difficulties to filter all incoming contents, to discover additional contents from outside of their friend circles, and importantly to share interesting contents with their different groups of interest.

For helping the users to overcome such difficulties, we propose a *User-centered and group-based approach for social data filtering and sharing*. This novel approach has a twofold purpose: (1) allow the users to aggregate their social data from different social network sites, and to extract from those data the contents of their interest, and (2) organize and share the contents within different groups. The members of a group are moreover able to choose which part of their social data to share with the group, and collectively define its topics of interest. To achieve the proposed approach, we define a modular system architecture including a number of extensible modules, and accordingly build a working Web-based prototype, called *SoCoSys*. The experimental results, obtained from the two different tests, confirm the added values of our approach.

Keywords: social media, social network sites, social data aggregation, information filtering, groups of interest, collaborative system

Résumé

Approche centrée utilisateur et basée groupe d'intérêt pour filtrer et partager des données sociales

par Xuan Truong Vu

Les médias sociaux occupent un rôle grandissant dans de nombreux domaines de notre vie quotidienne. Parmi d'autres, les réseaux sociaux tels que Facebook, Twitter, LinkedIn et Google+ dont la popularité a explosé ces dernières années, attirent des millions d'utilisateurs qui se communiquent, publient et partagent des informations et contenus à un rythme sans précédent. Outre les avantages reconnus, les réseaux sociaux ont également soulevé des problèmes divers. Nous sommes particulièrement intéressés par deux problèmes spécifiques : surcharge d'information et cloisonnement de données. Ces deux problèmes empêchent les utilisateurs d'exploiter pleinement et efficacement la richesse des informations poussées sur les réseaux sociaux. Les utilisateurs ont des difficultés pour filtrer tous les contenus reus, pour découvrir de nouveaux contenus au-delà de leurs réseaux personnels, et surtout pour partager les contenus intéressants avec leurs différents groupes d'intérêt.

Pour aider les utilisateurs à surmonter ces difficultés, nous proposons une *Approche centrée sur utilisateur et basée groupe pour filtrer et partager des données sociales*. Cette nouvelle approche a un double objectif : (1) permettre aux utilisateurs d'agréger leurs données sociales en provenance de différents réseaux sociaux, d'en extraire des contenus de leur intérêt, et (2) organiser et partager les contenus au sein de différents groupes. Les membres d'un groupe sont en outre en mesure de choisir quelle partie de leurs données à partager avec le groupe et définir collectivement les sujets d'intérêt de ce dernier. Pour implémenter l'approche proposée, nous spécifions une architecture de système comprenant plusieurs modules extensibles, ainsi que développons un prototype fonctionnel basé Web, appelé *SoCoSys*. Les résultats expérimentaux, obtenus des deux tests différents, valident les valeurs ajoutées de notre approche.

Mots-clés : médias sociaux, réseaux sociaux en-ligne, aggrégation de données sociales, extraction d'information, groupes d'intérêt, système collaboratif

Remerciements

Je tiens tout d'abord à remercier grandement mes deux directeurs de thèse, Mme Marie-Hélène Abel et M. Pierre Morizet-Mahoudeaux, pour m'avoir fait confiance en me proposant cette thèse, puis pour m'avoir guidé, soutenu, conseillé tout en me laissant une grande liberté pour proposer et poursuivre mes idées.

J'aimerais remercier Mme Kathia Oliviera, Mme Catherine Faron Zucker, M. Pascal Molli et M. Jean-Paul Barthes pour m'avoir fait l'honneur de participer au jury de soutenance, aussi pour leurs nombreuses suggestions et remarques pour améliorer la qualité du mémoire.

Je remercie particulièrement Nicolas, Vincent, Thibaut et toute l'équipe 50A qui m'ont accueilli parmi eux avec beaucoup de chaleur, de générosité, et aussi d'humour. Merci pour avoir rendu cette thèse beaucoup plus intéressante et enrichissante.

Mes remerciements vont également aux professeurs, aux thésards du laboratoire Heudyasic, et particulièrement de l'équipe ICI pour les échanges et les discussions enrichissantes non seulement sur les aspects scientifiques mais aussi sur les autres sujets d'intérêt. J'exprime aussi ma gratitude au personel administratif d'Heudiasyc et d'Uteam.

Merci à mes très chers amis Tien, Mi, Son, Marcio, Ana pour m'avoir accompagné, aidé pendant ces trois années de thèse. Merci à mes nombreux amis vietnamiens, français, brésiliens, espagnols, italiens, roumains, tunisiens, chinois et ceux venant d'autres horizons pour m'avoir montré que l'amitié est importante, précieuse et sans frontière.

Enfin, j'adresse toute mon affectation à ma famille, à mon père, à ma mère et à ma petite sœur. Merci pour leur confiance, leur amour et leur soutien sans faille pour moi malgré notre éloignement.

Je remercie à tous ceux qui ont contribué, de façon directe ou indirecte, à la réalisation et l'aboutissement de cette thèse. Merci à tous !

Contents

List of Figures

List of Tables

Abbreviations

API	**A**pplication **P**rogramming **I**nterface
BM	**B**oolean **M**odel
DOAC	**D**escription **O**f **A** **C**areer
FOAF	**F**riend **O**f **A** **F**riend
GUMO	**G**eneral **U**ser **M**odel **O**ntology
HTML	**H**yper**T**ext **M**arkup **L**anguage
HTTP	**H**yper**T**ext **T**ransfer **P**rotocol
IDF	**I**nverse **D**ocument **F**requency
IF	**I**nformation **F**iltering
IR	**I**nformation **R**etrieval
JSON	**J**ava**S**cript **O**bject **N**otation
LDA	**L**atent **D**irichlet **A**llocation
NLP	**N**atural **L**anguage **P**rocessing
N/A	**N**ot **A**pplicable
OPM	**O**pen **P**rovenance **M**odel
R&D	**R**esearch and **D**evelopment
RDF	**R**esource **D**escription **F**ramework
SCOT	**S**ocial **S**emantic **C**loud **O**f **T**ags
SIS	**S**ocial **I**nternetworking **S**ystem
SIOC	**S**emantically **I**nterlinked **O**nline **C**ommunities
SNA	**S**ocial **N**etwork **A**nalysis
SNS	**S**ocial **N**etwork **S**ite
SPARQL	**S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage
SQL	**S**tructured **Q**uery **L**anguage
TF	**T**erm **F**requency
UGC	**U**ser **G**enerated **C**ontent
UI	**U**ser **I**nterface
UML	**U**nified **M**odeling **L**anguage
URI	**U**niform **R**esource **I**dentifier
URL	**U**niform **R**esource **L**ocator
UTC	**U**niversity of **T**echnology of **C**ompiègne
VSM	**V**ector **S**pace **M**odel
WI	**W**eighted **I**nterests

This thesis is dedicated to my parents,

Chapter 1

Introduction

"The PC has improved the world in just about every area you can think of. Amazing developments in communications, collaboration and efficiencies. New kinds of entertainment and social media. Access to information and the ability to give a voice people who would never have been heard."

Bill Gates

Contents

The work presented in this thesis is funded by the Paris-based agency 50A[1] and is a part of its research and development (R&D) program, called "iBrain"[2]. iBrain was launched in February 2011, and carried out by the R&D team of 50A made up of web experts, developers and social media specialists. In March 2012, iBrain took a step forward by collaborating with the research group on *Information Knowledge Interaction*[3] of the UMR CNRS 7253 Heudiasyc laboratory of the University of Technology of Compiègne (France).

iBrain focuses on social media, in particular, on the plethora of social data generated every day by users. Its objectives are to design models and methods able to operate over different social media services in order to gather, merge, process data and to transform them into information and knowledge. These models and methods are furthermore intended to be combined and integrated within a single system dedicated to the management of social data. Thus, this project, first and foremost, called for learning about today social media, their diversity, their advantages and their challenges.

To begin with, in this chapter, we give an overview of the current state of social media. First, we will see some definitions of social media, their categories, and most importantly, their growing influence in many fields of our society. Then, we will dig a little deeper into social network sites which currently constitute the most representative category of social media. Some of the most successful social network sites will be thereby introduced. Next, we will discuss the existing issues and challenges raised by today social network sites. Especially, we will deepen into two particular problems, *information overload* and "*walled gardens*" which concern this work. Finally, we will reformulate these two problems in terms of a research question.

1.1   Social Media Landscape

Nowadays, social media have become a very important part of our every day life. People heavily and loosely use the term "social media" to refer to a wide range of online services including Facebook, LikedIn, Twitter, Youtube, Flickr, del.icio.us, etc. But, what is "social media"? what can we put under this term? and why are they so popular? This section attempts to answer the questions.

---

[1]Agence 50A: http://www.50a.fr
[2]Project iBrain by 50A: http://ibrain.fr/
[3]Research group ICI: http://www.hds.utc.fr/heudiasyc/recherche/equipe-ici/

1.1.1 Social Media Definition

"Social media" is one of the buzzwords that came along the advent of Web 2.0, somewhere around 2005 [112]. Many seem to use the two terms interchangeably, but it is worth noting that Web 2.0 is not a synonym for social media. It is a loose concept in reference to online services and technologies that give users who mainly played the role of consumer before, the possibility to become contributors as well. While Web 2.0 provides a functional environment for the realisation of social media [79], social media have become the central component of Web 2.0 [16].

Kaplan and Haenlein [81] defined social media as:

> "A group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of User-Generated Content (UGC)."

Accordingly, Kietzmann and al. [84] further specify that:

> "Social media employ mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, co-create, discuss, and modify UGC."

Basically, social media are composed of three core components, namely *people*, *community* and *UGC* (see Figure 1.1). It makes it possible for people to form online communities and share UGC [85]. The *people* may be the users of the open Internet or may be restricted to those who belong to a particular organization (e.g., corporation, university, professional company, etc.). The *community* may be a network of offline friends (whose friendship is extended to online), online acquaintances, or one or more interest groups (based on school attended, hobby, interest, cause, profession, ethnicity, gender, age group, etc.). The *User-Generated Content* may be created or brought from somewhere else by users and be of various types including photos, videos, bookmarks of Web pages, user profiles, user's activity updates, text (blog, microblog, and comments), etc.

Although current social media applications have a variety of features [84, 85, 93], there are generally 5 following typical characteristics [102]:

- *Participation*: encouraging voting, comments and sharing information, thoughts from everyone who is interested,

- *Openness*: being open to participation and feedback, removing protection boundaries,
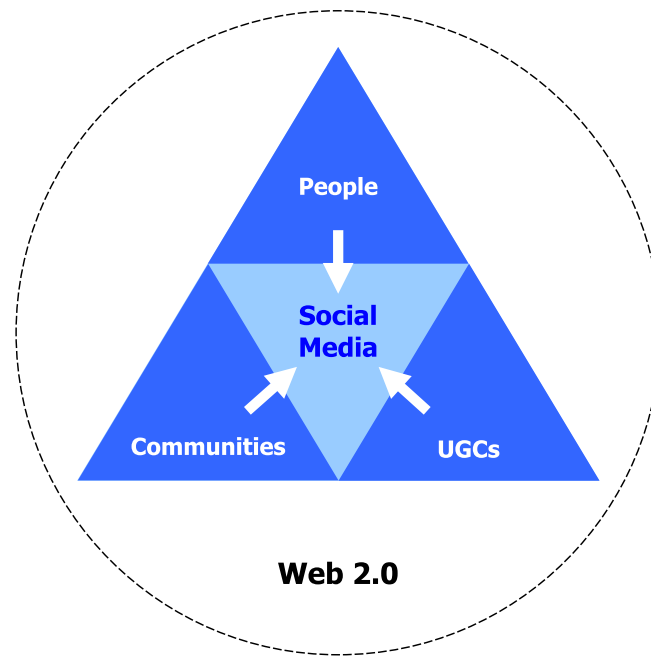
FIGURE 1.1: The core concepts of social media, adapted from those specified by [79]

- *Conversation*: fostering communications between users (i.e. one-to-one, many-to-many exchanges),

- *Community*: allowing communities to form quickly and communicate effectively,

- *Connectedness*: making use of links to other sites, resources and people.

1.1.2   Social Media Classification

A large number of social media applications exist through various forms and can be classified into multiple ways. Mayfield shown in [102] seven basic kinds of social media:

- *Social Network Sites*: these websites allow people to build personal profiles and then connect with friends to share content and communicate. The biggest social networks are, Facebook[4], Google+[5] and LinkedIn[6].

- *Blogs*: perhaps the best known and the earliest form of social media, blogs are online journals, with entries appearing with the most recent first. Blogger[7], Tumblr[8] and Wordpress[9] are some popular blogging services used today.

---

[4]Facebook: https://www.facebook.com/
[5]Google+: https://plus.google.com/
[6]LinkedIn: https://www.linkedin.com/
[7]Blooger: https://www.blogger.com/
[8]Tumblr: https://tumblr.com/
[9]Wordpress: https://wordpress.com/

- *Wikis*: these websites allow people to add content to or edit the information already available, acting as a communal document or database. The best-known wiki is Wikipedia[10], the online encyclopedia which has over 2.5 million English language articles [152].

- *Podcasts*: audio and video files that are available by subscription, through services like Apple iTunes[11].

- *Forums*: areas for online discussion, often around specific topics and interests. Forums came even before the term "social media" and are a powerful and popular element of online communities.

- *Content Communities*: communities which organise and share particular kinds of content. The most popular content communities tend to form around photos like Flickr[12], bookmarked links like del.icio.us[13] and videos like YouTube[14].

- *Micro-blogging*: social networking combined with bite-sized blogging, where small amounts of content such as short sentences, individual images, or video links are distributed online and through the mobile phone network in real-time. Twitter[15] is the clear leader in this field.

Kaplan and Haenlein [81] furthermore added *virtual worlds* as another kind of social media that replicate a three-dimensional environment, in which users can appear in the form of personalized avatars and interact with each other as they would do in real life. There are two forms of virtual worlds: virtual game worlds like World of Warcraft and virtual social worlds like Second Life[16].

It is important to note that these mentioned categories only represent the most common and mainstream part of the whole social media landscape. Many specified sub-categories and emergent categories of social media such as enterprise social networks, question-answer services, social commerce sites or location-based social applications, although they attract less users, also exist (See more categories in Appendix The Conversation Prism appendix).

Moreover, social media are a changing complex ecosystem as new services are created when other disappear and most evolve constantly. The distinctions among the different categories of social media are getting blurred. For example, social network sites and content communities overlap more and more. Whearas social network sites are adding primary

---

[10]Wikipedia: http://www.wikipedia.org/
[11]Itunes: http://www.apple.com/itunes/
[12]Flickr: https://www.flickr.com/
[13]Delicious: https://delicious.com/
[14]Youtube: https://www.youtube.com/
[15]Twitter: https://twitter.com/
[16]Secondlife: http://secondlife.com/

features of communities features, that is, the publishing and sharing of content, content communities are adding primary features of social network sites, that is, building personal profiles and forming communities.

1.1.3   Social Media Influence

Social media, over the past years, have rapidly increased in popularity and become a global phenomenon. They have been adopted by a wide range of demographic groups. Younger adults are especially enthusiastic adopters, but adoption rates for older people have grown as well. 72% of all Internet users are now regular users of social media and claim social networking their top online activity [31]. Therefore, they spend more time than ever on social media sites [53]. Facebook users spend approximately 20,000 years online and share 2.5 billion pieces of content each day. Also, 500 years of video and 400 millions tweets are posted on Youtube and Twitter per day [157], respectively.

There are many different reasons for the growing success of social media. Firstly, social media come in a wide variety of forms which are able to meet users' multiple needs. Secondly, most of them, in particular highly popular websites like Facebook and Twitter, are free, which makes them open to everyone. Thirdly, they are built in a way that makes them very user-friendly. Most sites are very easy to navigate and require little knowledge of the Internet. They are furthermore quite scalable as allowing a huge number of users to connect at the same time. Finally, social media have become ubiquitously accessible so that users can be connected from anywhere at anytime, thanks to the growth of mobile devices, in particular smartphones and tablets[17].

Social media have transformed the socio-cultural landscape - people's behaviours, attitudes, interactions and relationships [137]. They have introduced substantial and pervasive changes to communication between organizations, communities and individuals [84]. It becomes very easy to link to others, to create groups, and to form communities. The distance is cleared, cross border collaboration is then much facilitated. Constant flows of information and real time communication are enabled and allow people to spread their ideas, opinions, and thoughts at a great speed. Social media use is now widespread, mainstream and influential than ever. Individuals, organizations, companies, schools, and even governments are utilizing social media on a regular basis. Obviously, each of them have their own purposes and objectives.

---

[17]According to the Adobe 2013 Mobile Consumer Survey[10], 71% of surveyed people use their mobile device to access social media

1.1.3.1 Personal purposes

Many studies were carried out for understanding why people use and keep using social media [30, 39, 85, 96]. Reasons are multiple ranging from finding new friends to time killing. Essentially, there are three main dimensions of use [26]: the *Friendship Dimension*, the *Connection Dimension*, and the *Information Dimension*. The friendship dimension is about sustaining friend network for keeping in touch with both old and new friends and for locating old friends. The connection dimension is related to finding and making connections with new friends or like-minded people, and to feeling connected in general. The information dimension refers to the activities of gathering and sharing information with friends. Users may share information about themselves, post or look at pictures, learn about news, events. They may also engage in a one-to-one or one-to-many dialogs with their online network in order to seek or give answers or recommendations to specific questions of importance to them.

1.1.3.2 Professional and business purposes

Social media use for professional purposes is readily growing. Companies increasingly use platforms such as Twitter, LinkedIn and Facebook for recruitment [19], as they can promote their offers to a bigger number of potential candidates with less time. Job seekers are also turning to social media as their primary tool for job searching. More and more hiring managers and recruiters check candidates' social profiles, even if they are not provided, to study candidates' social behaviours. Thus, candidates may be rejected if inappropriate postings such as drinking, using drugs, negative comments about a former employer, lies about qualification, and so on are found on their profiles. Some employers may do similar things with their employees. They may decide to take disciplinary measures or even fire the employees, if the employees, even in off hours, are found to be in violation of the employer's code of conduct and confidentiality rules [85].

Beside, social media has become an efficient target of publicity and marketing [98]. Social media allow firms to engage (i.e. disseminate information, and receive feedbacks) in a timely and direct way with their end-consumers at relatively low cost and high efficiency [81]. That is not only relevant for large multinational firms, but also for small and medium sized companies, and even non-profit and governmental agencies desiring to reach and interact with their respective audiences [64, 97, 107]. However, using social media is not an easy task and requires new ways of thinking. Brands need to have contents prepared professionally. They also need to assign qualified employees to manage their presence on social Web sites to deal with user comments and requests, gauge the tenor of reactions from the users, etc. Negative reactions can spread widely and quickly on social media [85].

1.1.3.3   Political campaigns

Social media is increasingly used for political and civic purposes. People are likely to post their own thoughts about issues, post links to political material, encourage others to take political action, belong to a political group on a social networking site, follow elected officials on social media, and like or promote political material that others have posted. On the other hand, politicians can tap into a wealth of information about the people who are following them on social media, and customize their messages based on selected demographics. They can also reply on social media analytics in order to weight public opinion on their policy statements and moves during their campaigns.

The election of Barack Obama as President of the United States in 2008 [138] and the so-called "Arab Spring" in the Middle East in early 2011 [62] are, among others, two strong examples of the political power of social media. In both events, social media played a decisive role, as it helped organize such demonstrations, mobilize their activists. Most importantly, it allows a constant update and dissemination of news of the events locally and globally.

1.1.3.4   Social mobilization

The use of social media for responding to emergency events, in particular natural disasters, and creating situational awareness has risen in recent years [56, 145]. In many disaster situations, people post situation-sensitive information on social media related to what they experience, witness, and/or hear from other sources that allows both affected populations and those outside the impact zone to learn about the situation, the state of their homes and families [74].

Many emergency responders and humanitarian officials recognize the value of the information posted on social media platforms by members of the public (and others), and are interested in finding ways to quickly and easily locate and organize that information that is of most use to them [73]. For example, social media can be used to solicit support for resources from people to aid affected victims or can be utilized as a means of publicizing the picture, names and addresses of missing persons so that relatives, friends or anyone that finds them can easily help with reuniting them to their loved ones.

1.1.3.5   Corporate purposes

The wide acceptance of social media by the public has led numerous organizations to search for using social media inside organizations. Many studies [44, 52, 144] have explored ways

that would help organizations to take the opportunities to improve their organizational effectiveness and to increase their employees' work performance. Social media, especially social network sites, typically allow the management to post announcements to employees, encourage employees to learn more about each other, to share work-related documents, materials and exchange messages. That gradually transforms organizations' hierarchical structure to some "networked" structure, which is more conductive to internal coordination, knowledge sharing and teamwork.

Using public social network sites at work may cause serious problems such as business secret leaking and personal information disclosure. Organizations therefore need "closed" services. There are two ways in which organizations can operate closed social network sites. One is to create a closed enterprise social network on an open social network site (i.e. private group on Facebook, or company group on LinkedIn). Another way is to create a closed enterprise social network by using enterprise social networking software, such as IBM Lotus Connections, or Microsoft Sharepoint. In the last case, the associated data relating to the connections, interests and activities of employees are made available by the organization, providing new information sources and new possibilities to get meaningful insights aiding to understanding the internal communication, and even to making decisions.

To sum up, it is important to realise that social media play a very important role in many areas of our society. From a general point of view, social media give users two major features, *information* and *conversation*. These two features offer the public an unprecedented power. People can now compete with traditional media by publishing, communicating, and sharing their own information, opinions and thoughts. Moreover, conversation barriers, both distance and hierarchy, have been removed, as anyone can talk in a direct and timely way to anyone including individuals, companies, organizations, and governmental agencies.

1.2 Major Social Network Sites

Social network sites (SNSs), also called *social networking sites* [81, 85], *online social networks* [69], or *social network services* [87, 88], are websites whose main goal is to congregate and to connect people. They allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system [29]. SNSs also enable users to invite friends and colleagues to have access to their profiles, and to send e-mails and instant messages to each other [81]. SNSs have constantly evolved to comply with users' upcoming demands. These originally profile-centric platforms recently facilitate and encourage publishing, sharing information and contents as well.

Over the last several years, SNSs have exploded in popularity. They are now occupying the central place of the ecosystem of social media. Facebook, Twitter, LinkedIn or Google+ websites are the most successful examples. Each of them claims hundreds of millions of active users worldwide [135] and belong to the ten most-visited websites on the Web [13]. Although sharing aforementioned common features, they have their own features and operations that we examine below.

1.2.1   Facebook

Facebook was founded on February 4, 2004 by Mark Zuckerberg and initially used as a Havard-only social network. It quickly expanded to other schools then to high schools, businesses and finally to anyone who claims to be at least 13 years old by 2006. In 2008, Facebook became the most popular social network site. Today, Facebook is accessible in almost every country and through 78 languages, and is used by 1.23 billion *monthly active users*[18] worldwide. 88% of users are said to be loccated outside US and Canada [50].

After registering to Facebook, users need first to set up a personal profile. Users' Facebook profiles are currently very elaborate covering not only demographic information but also information about schools, works, interests, and so on. Users may add other users as friends and then begin to exchange messages and post status messages, photos, videos or links.

Facebook has constantly developed, experienced and rolled out a number of features including News feed, Privacy. News feed appears on every user's homepage and highlights information including profile changes, upcoming events, and birthdays of the user's friends in a chronological order so that the user can comment, like or share information rapidly. To allay concerns about privacy, Facebook enables users to choose their own privacy settings and choose who can see specific parts of their profile and their posts as well (e.g. only friends, friends of friends, everyone). Additionally, Facebook allows users to create, to invite others or to join groups and events. Facebook attracts not only individuals but also companies and organizations for professional and business purposes. Being on Facebook allows them to reach, engage and interact directly with their consumers.

Facebook moreover opens its user database, called *social graph*, to third-party applications and websites via its APIs, that makes Facebook more of a platform than a single service where users' social experience is definitively extended. In brief, Facebook is a general-purpose social network site where people are mainly connected for keeping in touch with family and friends.

---

[18]*Monthly active users* is a metric counting the number of unique users per the past 30 days

1.2.2 Twitter

Twitter was launched in July 2006 as the first *microblogging* service combining social networking and micro-blog publishing. The service then rapidly gained worldwide popularity, in particular in North America and Japan. To date, Twitter supports more than 35 languages with 271 million monthly active users worldwide who send nearly 500 million tweets per day [148].

Twitter enables users to send and read short messages within 140 characters called *tweets*. This unique feature makes it easy for anyone to quickly create, distribute and discover content, and subsequently drives a very high information exchange rate that makes Twitter highly "live". Twitter furthermore introduced *hashtags*, words or phrases prefixed with a # sign. Using hashtags allows users to efficiently group posts together and rapidly search them by topic.

Unlike other social network sites which require reciprocal relationship, Twitter relies on a *following-follower* relationship base. Any user can follow any other member (i.e. *following*) to receive tweets from them as well as be followed by any member (i.e. *follower*) who desires to receive his/her tweets. Tweets can also be *retweeted* by others for their respective network of followers to enhance the audience. In short, Twitter is a blend of instant messaging, blogging, and texting, but with short content and a very broad audience.

1.2.3 LinkedIn

LinkedIn was launched in May 2003 and was one of the first social network sites devoted to the business community. The website has started to gain popularity since late 2006. In April 2014, LinkedIn reached 300 million registered members in more than 200 countries and territories. 67% of LinkedIn members are located outside the United States [95].

LinkedIn allows users to create profiles, basically resumes with emphasis on employment history and education, and *connections* to each other. Unlike other free social network sites, LinkedIn requires connections to have a pre-existing relationship. A member with basic membership can only connect with someone that he/she has worked with, knows professionally or has gone to school with. He/she is not allowed to contact other users through LinkedIn without an introduction or a recommendation from LinkedIn. Moreover, connections can only interact through private messaging.

The main functionality of LinkedIn is to link employees, employers, and companies with each other. Employees can look for jobs, people and business opportunities whereas employers can list jobs and search for suitable candidates and companies can find potential

clients or service providers. Each user can furthermore introduce or recommend someone in their contact network to new available offers. Gradually, other features have been also added, such as groups, question and answer forums for exchanging knowledge and expertise. Briefly, LinkedIn is a professional network site mainly used for searching and advertising for jobs or business opportunities.

1.2.4   Google+

Google+ was launched in June 2011 and was considered as Google's biggest attempt to rival the social network Facebook [35]. Google+ saw an explosive growth during its early time, as it only took three months to reach some 50 millions users (years for other social networks) [46]. It is now the second-largest social network site in the world after Facebook with 359 million monthly active members worldwide [43].

Google+ is quite similar to Facebook with a lot of comparable features (e.g. *stream* versus *news feed*, *circles* versus *lists*, or *+1 button* versus *like button*). There are however some clear differences. For example, the feature *circles* allows a more customizable classification of one's contacts (family, friends, colleagues, and others) and allows users to choose to which circles of users they want to show their content. Furthermore, like Twitter, Google+ does not require reciprocal relationships between users, but a unidirectional consent. Most importantly, Google+ is intentionally connected to other popular Google Web services such as Gmail, Youtube, Google Hangout or even Google's search engine so that users can seamlessly use these services. In short, Google+ is the *social layer* of the entire ecosystem of Google.

Facebook, Twitter, LinkedIn and Google+ constantly try to differ from each other by regularly adding new and specific features allowing users to even better personalize their social experience. Users therefore tend to use many or most of these websites in order to take full advantage of various features provided by each website. Such a significant membership overlap along with a huge number of users make these four social network sites very representative and powerful datasets for studies and researches [8, 67, 78, 125]. Likewise, instead of trying to consider the entire social media ecosystem, which, as described earlier, is extremely large and various, we concentrated our efforts in studying SNSs, in particular these four social networks.

1.3   Social Network Problems

The growth of SNSs, in particular large-scale websites like Facebook, Twitter, LinkedIn, and Google+, in terms of the number of users, the level of daily traffic, and the amount of

UGC, has been absolutely incredible. These particular social media services now have an enormous impact, both positive and negative, in our society. Alongside the aforementioned benefits, they have raised various issues and challenges [69, 85] as well. We categorized the existing problems into two main families. The first family includes problems associated with the users' misuses of SNSs. The second family includes problems related to the inherent characteristics of current SNSs.

1.3.1 User Misuses

User misuse issues are those that misuses of SNSs have brought about, whether intentionally or unintentionally, by some users and impacted on themselves and/or other users.

One of the major problems is SNS addiction [86]. Many users seem to lose their appropriate sense and spend excessive amounts of time to post and view every small and trivial updates of their activities and their so-called online friends' activities. Some furthermore suffer from reduced productivity by spending time on SNSs while at work.

Another major problem refers to the privacy of SNS. Some users indiscreetly disclose too much details about their identity, their activities, and thoughts on SNSs [156]. It can reveal something about users that they would rather want their current or future employers or school administrators not know. Additionally, inappropriate postings may lead to some legal issues.

Security is another critical aspect of SNSs [42]. Pedophiles/sex offenders and terrorist groups join popular SNSs to hunt for their preys and recruit their new members respectively. Many spammer, phishing and malware attacks targeting unsuspecting users have been also detected on SNSs. All of this can put your personal safety at risk.

Besides, damages can be also caused by irresponsible users. Some instigate or participate in cyber bullying and cyber stalking that may lead to occasional suicides. Others spread false rumours or information which, in the absence of an efficient verification mechanism, may heavily damage someone's reputation and business.

These negative aspects are only some illustrative examples of issues associated with the users' misuses of SNSs. Given the high velocity and the breadth of today SNSs, it is very challenging to cope with such problems. They deserve many efforts including personal, social and governmental efforts as well as the willingness of the SNS providers.

1.3.2   Inherent Problems

The inherent problems are originated from the specific characteristics of current SNSs.
They do not alter the users' behaviours, but prevent them from efficiently exploiting SNSs.
Information overload and "*Walled gardens*" are two major problems of this family. The
first problem is caused by the continued growth of SNSs, whereas the second problem is
due to the fact that the current SNSs are proprietary and disconnected from each other.
We further analyse both problems.

1.3.2.1   Information Overload Problem

Information overload is a general problem that occurs in a wide variety of disciplines [47].
Information overload occurs when the amount of input to a system exceeds its processing
capacity [141]. Individuals have fairly limited cognitive processing capacity. Increasing
information at first increases an individual's capacity but eventually additional information
becomes unhelpful and information-processing ability declines [27].

SNS users are increasingly facing this problem of information overload [41, 63, 121, 160],
as they are often overwhelmed by the huge number of incoming contents. On SNS, users
typically receive contents from their social friends and other accounts representative of
organizations, companies that users follow. As most of SNSs do not tightly restrain the
number of ties that a user can add as friend or follow, many users' social networks are
actually very large. For example, on Facebook, an average user has 338 friends [1] and
follows 40 pages[19] [89]. The same user may, via these social connections, receive per day
hundreds of various pieces of content including profile updates, posts, photos, videos, links,
tags, check-ins, and so forth. This is much beyond the time that the user can devote to
process all contents. Certain SNSs provide features such as keyword/hashtag search as a
naive solution for the information overload problem, but these filters are not sufficient to
provide complete personalized information for a user.

On the other hand, popular SNSs like Facebook and Twitter respectively implemented
features like *News Feed* and *Tweets Timeline* which appear on every user's homepage, and
display current updates and activities of their social friends into a single stream so that
users can easily and rapidly react, like, or comment. Such features are useful, as users may
not want to browse all of their friend list one by one, each time they are connected to the
SNS. However, there is a major drawback with this streamlined presentation when it comes
to searching for contents of interest. Indeed, contents are generally shown in the order of

---

[19]Facebook pages are for businesses, brands and organizations to share their stories and connect with
people. Like profiles, you can customize Pages by posting stories, hosting events, adding apps and more.
People who like your Page and their friends can get updates in News Feed.

their timestamps and in proportion to the activity level of users' friends, regardless of their topics. Users therefore have to go through the stream to locate information that likely interest them, among lots of other undesired information.

Some SNSs allow users to organize their friends into smaller lists according to users' own criteria (e.g. "Colleagues", "TV show comedians", "Hightech related people", etc.) so that they can follow these connections separately. This listing task is however not easy and could even lead to unsatisfactory results because of two clear reasons. Firstly, users do not share only interesting contents but a lot of personal stuff as well [110]. Secondly, one's interests may change over time that requires users to maintain and adjust their lists regularly.

Consequently, many important and interesting pieces of information remain unnoticed by users, whereas lots of irrelevant and contents not worth reading keep showing up.

1.3.2.2 Walled-Garden Problem

As mentioned above, it is common that one user engages with multiple SNSs in order to take advantage of various free features provided by each website. According to a survey conducted by Princeton Survey Research Associates International in 2013[20], 42% of those interviewed claim use two or more of 5 SNSs (i.e. Facebook, Twitter, LinkedIn, Instagram[21], and Pinterest[22]) [45]. Especially, there is a significant level of overlap between Facebook, Twitter and LinkedIn users (See Table 1.1).

TABLE 1.1: **Social Media Matrix** - Pew Research Center (2013)
% of users of each particular site who use another particular site
(e.g. 90% Twitter users also use Facebook)

	Use Facebook	Use Twitter	Use LinkedIn
% of Facebook users who...	N/A	22	25
% of Twitter users who...	90	N/A	39
% of LinkedIn users who...	83	31	N/A

These SNSs and others are however centralized. The companies providing the services have the sole authority to control all the data of the users [155]. The identity of a user and their data can easily be entered, but only accessed and manipulated via proprietary interfaces, so creating a "wall" around connections and personal data [14], as illustrated in Figure 1.2.

---

[20]The survey was conducted from August 7 to September 16, 2013, among a sample of 1,801 adults, age 18 and older. Interviews were conducted in English and in Spanish, and on landline and cell phones.

[21]Instagram: http://instagram.com/

[22]Pinterest: http://www.pinterest.com/

The SNS providers just do not want their users to be active on other websites, given that every user is pure capital to them.

The lack of interoperability across SNSs leads to many problems such as portability, identity, linkability, and privacy [14]. Most importantly, that makes it difficult for users to transform, reuse data including their profiles, their social networks, the messages with their friends and their photos among different SNSs. Every time a user goes to a new site, he/she has to create a new profile, re-enter his personal information, connect again to his/her friends, and so forth. Thereby, users' activities and their friendships are scattered across different SNSs. It becomes increasingly inconvenient for users to manage their social data and constantly check several SNSs to keep track of all recent updates [160]. As a result, some users have ended up by reducing, or even stopping their activities on certain SNSs in order to focus on others.
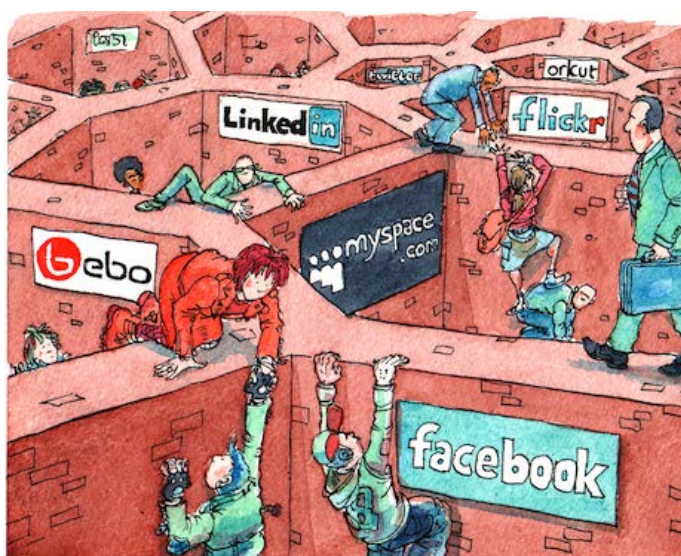
FIGURE 1.2:   Social Networking Sites as Walled Gardens by David Simonds, The Economist, 19 March 2008

Disconnected SNSs moreover hinder the information sharing between users. There is no direct way for a user of a given SNS to send a piece of information to other users of another SNS. Thus, the user needs to be member of both SNSs and has to duplicate that same information on each SNS, if he wants his friends on both SNSs to receive it.

1.4   Research Question

Although the first family of user misuse problems are very important, they go far beyond a thesis work in the domain of computer science, and requires multi-discipline approaches and solutions. We will mainly address the two aforementioned inherent problems, *Information*

overload and "*Walled gardens*". These two problems raise a number of interesting questions which open new perspectives and call for researching suitable solutions. We focus on three of them which are *Filtering, Discovering* and *Sharing* questions that we detail below.

1.4.1 Filtering Question

The two problems, *Information overload* and "*Walled gardens*", make it very time-consuming and laborious for users to manually extract contents of interest, as they have to navigate social network by social network, and to go through all incoming information. Obviously, most users cannot spend such a considerable time and effort. Consequently, they are missing many important and interesting pieces of incoming information. So, the *Filtering* question can be stated as:

Q_1: *How to help users to extract contents of interest from their different social networks with less effort and without altering their social networking experience?*

This question is made of four distinct parts corresponding to the four major requirements to be fulfilled:

1. *To extract contents of interest,*

2. *From the users' different social networks,*

3. *With less effort,*

4. *Without altering the users' social networking experience.*

The first requirement narrows the set of possible solutions to the field of personalized information filtering. The second requirement extends the scope of the expected solution to multiple SNSs. The third requirement imposes that the expected solution has to reduce the users' manual effort to a minimum. The fourth requirement means that the users do not need to change anything in their use of SNSs, and keep interacting with their social networks normally.

1.4.2 Discovering Question

The SNS providers have established privacy rules restricting what a user may receive within his/her social streams. In general, only the contents shared by the user's social friends are displayed. Contents, in particular, contents of the user's interests, from outside the circle

of friends are hidden from the user's view. The user can still add new friends to expand his/her circle of friends, subsequently the number of information sources. However, this also increases the chance of information overload. Hence, the *Discovering* question can be stated as:

Q_2: *How to help the users discover additional contents of interest from outside of their circles of friends?*

This question also leads to a closely related question:

Q_2': *What are the information sources to be considered, given that the users' interests are various and changing?*

1.4.3   Sharing Question

People are often part of different groups of interest, which are held and driven by a common interest. It may be a hobby, something the group members are passionate about, a common goal, a common project, or merely the preference for a similar lifestyle, geographical location, or profession [154]. Taking part in the group enables its members to exchange information, to obtain answers to personal questions or problems, to improve their understanding of a subject, to share common passions or to play [70].

Groups of interests impose a group setting, which makes sure that the members share only contents related to one or several particular topics at a single place. This makes it much easier to discover interesting information and useful contents. Nevertheless, the group commitment degree is different among members. Often, it is only a small number of members who actively generate contents, while the majority of members are passive consumers. Therefore, a group may be short of good contents if its active members are no longer active. This is more and more common, as people get used to push interesting information on the different social networks to maintain their social presence and social influence [39] while forgetting to also share it with their interested groups. Moreover, there is no guarantee that the members of a given group are connected to the same social network sites, or connected to each other. Hence, interesting information published by a member on a particular social network is not necessarily visible to other members.

Some groups of interest may be formed within a particular SNS. For example, Facebook and LinkedIn provide features allowing users to create, join groups, ask and share specific

information inside the created groups. However, these groups are exclusive to the corresponding social networks (i.e. kept behind "walled gardens"). The interaction between two groups, for example a Facebook group and a LinkedIn group, is not directly possible, even though they share a common interest and many identical members. One member has to manually copy the same information to the two groups to ensure that it could be shown by everyone.

So far, there has not been an efficient solution for a group of interest to tap into the contents published by its members across different SNSs to retrieve some parts relevant to its topics of interest. As a result, the group is missing lots of interesting information. This raises the *Sharing* question:

Q_3: *How to make it possible for the users to share the contents of their various social streams with their respective groups without extra charge?*

The "*without extra charge*" part of the question is very important. It requires that the contents matching the group's topics of interest should not be selected and shared with the group manually by the users, but in an automated manner. That saves the users from extra manual efforts while making sure that the group receives regularly new and interesting contents.

1.5   Summary

Social media have become a very important part of our every day life. There are a large number of social media services that exist through various evolving forms. Among others, social network sites have exploded in popularity over the last few years. Facebook, Twitter, LinkedIn, and Google+ are some of the most successful examples with unprecedented number of users, daily traffic, and amount of generated data. Basically, these websites allow people to build personal profiles and then connect with friends to share content and communicate. Each of them furthermore offer users specific features. It is therefore very common that a single user is connected to most of or even all of these websites simultaneously.

We therefore concentrate our efforts in studying SNSs, especially the four cited websites. Alongside the benefits, SNSs have also raised various issues and challenges. Some problems are associated with the users' misuses of SNSs when others are originated from the inherent characteristics of the current SNSs. The second family, in particular the two detailed above problems *information overload* and "*walled gardens*", concern this work. They prevent users from efficiently exploiting the current SNSs. Users have difficulties to filter all

incoming information, to discover information from outside of their circles of friends, and importantly to share the interesting contents with their groups of interest. With respect to such difficulties faced by users, we will address the three corresponding questions: *filtering*, *discovering*, and *sharing*. They will serve as reference points for our whole work.

The rest of this thesis is organized as follows. In Chapter 2, we will review the works which have already addressed the three questions. As we will see, they propose only partial solutions, if at all, and cope with only one or two questions. They furthermore present some limitations to be directly applied to our case. In the following chapters, *Conceptual Design* and *Technical Solution*, we will present our adapted answer, its conceptual and technical aspects respectively. In Chapter 5, we will present a working Web-based prototype as proof of concept. We will discuss about the findings obtained from two testing experiences with real users using the same prototype in Chapter 6. These findings show some encouraging results confirming the added values of our work. Finally, before recalling the main contributions of this thesis in the Summary chapter, we will set out some perspectives for future work.

Chapter 2

Literature Review

Contents

Users constantly generate a myriad of data on SNSs. These social data are extremely rich, since they include not only personal data but also relational data. While personal data may unveil a lot about the involved individuals, for instance, their education, employment, interests, and so forth, relational data furthermore describe their social interactions in terms of how, when and with whom they share information [3].

Thus, SNSs contains a wealth of real-world and live information for researchers and practitioners in multiple disciplines. A significant number of researches and studies on SNSs

were reported in [20, 82, 87, 161]. Researchers have utilized SNSs, especially the generated
social data for sensing real-world events [17, 40, 126], for detecting key users [51, 150], or
for analysing online social networks [12, 25, 54, 158]. Likewise, enterprises and govern-
mental organizations, using suitable techniques for data gathering and content analysis,
have attempted to obtain from social data meaningful insights and knowledge about brand
exposure, brand community and acceptance by users [54, 58, 108].

Our bibliographical study allowed us to identify many interesting works. Most of them
tried to address the "walled gardens" problem (e.g. [7, 37, 105, 113]), or the information
overload problem (e.g. [48, 59, 129, 136, 143]), or both (e.g. [160]). Some others proposed
different recommendation methods with the objective to help users discover new interesting
contents (e.g. [6, 38, 80, 134]). We have not found any work studying the benefits and/or
proposing solutions for helping users to share the contents of their various social streams
with their respective groups.

It is worth noting that the identified works used various techniques from different research
fields such as *data portability*, *natural language processing*, *data mining*, *recommender sys-
tems*, *information retrieval*, and so forth. Given such interrelation with these domains, it
would be difficult to analyse the works domain per domain. Therefore, we prefer to review,
in this chapter, the related works with respect to our addressed questions and classify them
within three respective sections, *Social Network Aggregation*, *Information Filtering*, and
Information Discovering.

2.1   Social Network Aggregation

Social Network Aggregation is a common solution to the "walled gardens" problem of
current SNSs. It seeks to collect the various social data of a given user from different SNSs
into one unified presentation while attempting to organize and simplify the user's social
networking experience. Social Network Aggregation is basically a three-step process :

1. *Unique User Identification* : Identify the user's different accounts across SNSs,

2. *Data Collection* : Access and retrieve the user's various social data,

3. *Data Representation* : Define a common model for representing the heterogeneous
   social data.

Different alternatives have been proposed for each of these three steps. We will respectively
deepen them below. Then, we will show several representative commercial Social Network
Aggregation services.

2.1.1   Unique User Identification

Popular SNS such as Facebook, Twitter and LinkedIn did not use open and interoperable protocols like OpenID[1], and instead implemented their custom authentication management services mostly based on the OAuth framework[2], for example Facebook Connect, or Sign in with Twitter. Each time a user goes to a new site, he has to create a new profile, re-enter his personal information, connect again to his friends, and so forth. As a result, a user may have different social identities across SNSs.

Therefore, the first requirement when aggregating a user's social networks is to identify his/her various social identities. People search engines such as Peekyou[3], Pipl[4], or one introduced in [37], allow searching the different social accounts of a user, based on some public personal attributes, for example name, username, email or location. However, a user may set different values to these attributes, or even leave them undefined, which makes the identification incomplete.

Google proposed an alternative, named Social Graph API[5], which crawled users' personal web pages, essentially Google profiles, and extracted links referred to their other social profiles. Given a user's URI of an online account, the API would return all available mappings. Unfortunately, it has been withdrawn by Google.

Another straightforward way is to systematically implement for each SNS its corresponding authentication protocols and to ask users to directly authenticate their social identities.

2.1.2   Social Data Collection

After identifying a user's social accounts, the next step is to collect (i.e. access and retrieve) the social data associated with these social accounts. For that purpose, several methods were discussed in [153], of which there are two automated techniques : (1) scrapping the user's profile pages, and (2) using the APIs provided by the SNS providers.

The first technique consists of crawling the user's profile pages with an automated script that scans and extracts the wanted information from HTML codes using HTTP requests and responses. This approach does not require the implementation of specific protocols and the provision of authentication data [37]. However, it is only possible when the SNS

---

[1]OpenID is a single sign-on system that allows users to log on across multiple sites without having to register with their information over and over again. http://openid.net/

[2]OAuth : http://oauth.net/

[3]Peekyou : http://www.peekyou.com/

[4]Pipl : https://pipl.com/

[5]Google Social Graph API : https://developers.google.com/social-graph/

providers do not disallow it in their terms and conditions. Few providers open a very small set of public information for scrapping while many others totally prohibit this practice.

The second technique requires first to register an application with the SNS. Users then have to give suitable permissions to the application so that it can send relevant queries to the corresponding API of the SNS for collecting data. With this technique, an application is not limited to public information, but can access much more users' social data. Using proprietary APIs and specific authentication services provided by the SNS providers has become a common practice for most developers. There are however two considerable drawbacks. Firstly, the SNS providers often restrict the number of API calls that an application can make for a certain time interval. Secondly, provided features vary greatly from one API to another, which requires developers to learn how to handle each API.

Regarding the latter issue, OpenSocial[6] has been developed as a first attempt for standardization of APIs. In fact, it provides a common cross-platform API, which gives access to a number of supported SNS. More than 80 SNSs have currently subscribed to Google Open Social [106]. Nevertheless, many popular SNSs like Facebook and Twitter do not support it yet.

Another way for collecting users' social data is to reply on commercial solutions like GNIP[7], Datasift[8], or Topsy[9]. These data vendors are the premium partners of some major SNSs that allows them to access to full real-time streams of public data from these SNS. One important advantage is that data are already aggregated from different sources whenever the customers want to buy them.

2.1.3 Social Data Representation

SNSs use their own syntaxes and terms in order to describe users and their social activities. A same kind of information may be called differently across SNSs. Therefore, to be able to consolidate heterogeneous social data, a unified representation model is required. Such common data model is crucial, as it allows the integration of data gathered from various sources, instead of a mere juxtaposition.

Researchers have put a lot of efforts into developing many generic user models [36]. One remarkable effort is the General User Model Ontology (GUMO) [68]. This ontology is intended to cover all aspects of a user's life ranging from contact information and demographics over abilities, personality right up to special information like mood, nutrition or

---

[6]OpenSocial : http://opensocial.org/
[7]GNIP : http://gnip.com/
[8]Datasift : http://datasift.com/
[9]Topsy : http://topsy.com/

facial expressions [119]. Despite a large number of important dimensions, GUMO lacks properties relevant to SNS users such as social accounts or user interests.

TABLE 2.1: **Semantic Web Vocabularies** for representing social data

Ontology	Description	Purpose
FOAF[10]	Friend of a Friend	Describing persons, their activities and their relations to other people and objects
Relationship[11]	Relationship	Specializing the type of people relationship (e.g. familial, friendship or professional relationships)
DOAC[12]	Description Of A Career	Representing the past working experiences of the users and their cultural background
GeoNames[13]	GeoNames	Adding geospatial semantic information to user locations
SIOC[14]	Semantically-Interlinked Online Communities	Representing user activities in blogs and forums
SCOT[15]	Social Semantic Cloud of Tags	Describing user tagging activities
WI[16]	Weighted Interests Vocabulary	Representing user interests and their corresponding degrees
OPM[17]	Open Provenance Model	Stating that an interest was originated by a specific website

Another approach is to combine a number of light-weight ontologies which have already been widely adopted by the Semantic Web community to depict users and their activities on the Web. Some Semantic Web vocabularies useful for representing social data are listed in Table 2.1. Such approach is increasingly getting attention, as many researchers, in their related works [7, 57, 80, 105, 113, 124], showed that a large part of social data could be translated into the corresponding semantic counterparts provided by these vocabularies.

Additionally, some authors [57, 105, 113] proposed to use interlinked datasets on the Web of Data, such as DBpedia[18], in order to semantically enrich social data (see Listing 2.1). It requires hence an extra step of analysis that identifies entities from the text retrieved at the previous stage and links them to URIs, for example, on DBpedia. Available named entity

---

[10] FOAF project : http://www.foaf-project.org
[11] Relationship specification : http://vocab.org/relationship/
[12] DOAC specification : http://ramonantonio.net/doac/0.1/
[13] GeoNames specification : http://www.geonames.org/ontology
[14] SIOC project : http://www.sioc-project.org/
[15] SCOT specification : http://rdfs.org/scot/spec/
[16] WI specification : http://purl.org/ontology/wi/core#
[17] OPM Specification : http://openprovenance.org/
[18] DBpedia is the semantic representation of Wikipedia. It is currently the largest cross-domain dataset on the Web of Data. http://dbpedia.org/About
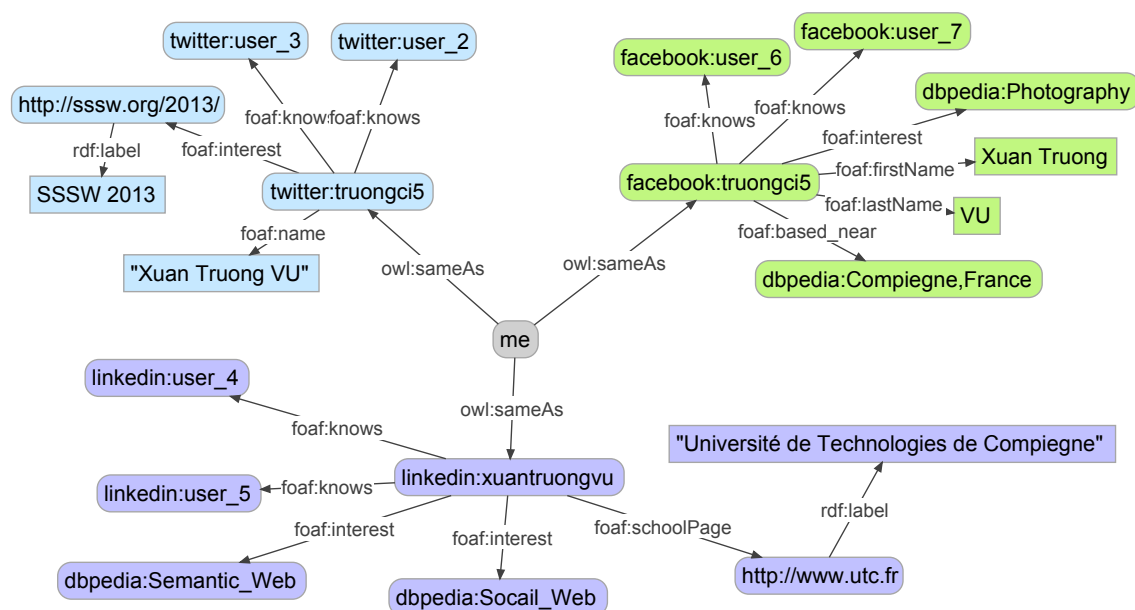
FIGURE 2.1: Example of merging a user's different social profiles using `owl:sameAs`

recognition services like Zemanta[19], DBpedia Spotlight[20], Alchemy API[21] can be used to perform such a task [123].

```
@prefix sioc: <http://rdfs.org/sioc/spec/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix dbpedia: <http://dbpedia.org/resource/> .

<http://twitter.com/bob/status/73748435752333312>
    a <sioc:Post> ;
    dcterms:created "2011-05-26T15:52:51+00:00" ;
    sioc:has_creator <http://twitter.com/bob> ;
    sioc:content "Awesome, love the new Garageband for iPad http://is.gd/
    SJqVav" ;
    sioc:links_to <http://is.gd/SJqVav> ;
    sioc:has_topic dbpedia:Apple_Inc.  ;
    sioc:has_topic dbpedia:GarageBand ;
    sioc:has_topic dbpedia:IPad .
```

LISTING 2.1: Example of representing a user's post using SIOC and DBpedia

Beside, there are different proposed techniques for merging a user's social data from different SNSs. Bojars et al. [24] suggested to use two semantic properties, namely `owl:sameAs` and `rdfs:seeAlso` to associate a user with his/her existing social profiles as illustrated

---

[19]Zemanta : http://www.zemanta.com
[20]DBpedia Spotlight : http://dbpedia.org/spotlight
[21]Alchemy API : http://www.alchemyapi.com

in Figure 2.1. Abel, Gao et al. [7, 57] simply put all aggregated information under one unique entity. Some user attributes could therefore get several identical or different values extracted from various profiles. Kapanipathi and Orlandi et al. [80, 113] inserted the provenance information into each user interest by means of the Open Provenance Model (OPM) (see Listing 2.2).

```
@prefix wi: <http://purl.org/ontology/wi/core#> .
@prefix wo: <http://purl.org/ontology/wo/core#> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix ex: <http://example.org/stuff/1.0/> .
@prefix opm: <http://openprovenance.org/model/opmo#> .

ex:me
    wi:preference [
        a wi:WeightedInterest ;
        wi:topic dbpedia:Semantic_Web ;
        wo:weight [
            a wo:Weight ;
            wo:weight_value 0.5 ;
            wo:scale ex:Scale ] ;
        opm:wasDerivedFrom <http://twitter.com/me>
        ] ;
```

LISTING 2.2: Example of using OPM to state the provenance of a piece of information, in this case, a user interest

Another important work of standardization is the Activity Stream project[22], which is an effort to develop a protocol to syndicate activities taken by users in different SNSs. In its simplest form, an activity consists of an *actor*, a *verb*, an *object*, and an optional *target*. It basically tells the story of a person performing an action on or with an object, for example "Bob posted a photo" or "Bob liked a video of a friend". One important advantage of Activity Stream is that its wide range of verbs and objects[23] were directly inspired from users' real activities on current SNS, thus reflecting very well user social networking experience. Another advantage is that many SNSs including Facebook have already implemented their user activity streams with Activity Stream and opened up them to developers to use.

2.1.4 Social Network Aggregators

Services implementing Social Network Aggregation are commonly called *Social Network Aggregators*. These services allow a user to consolidate at a single point the various social activities in such a way that the user is not required to login to each SNS and perform

---

[22]Activity Stream project : http://www.activitystrea.ms/
[23]Activity Streams - Base Schema : https://github.com/activitystreams/activity-schema/blob/master/activity-schema.md
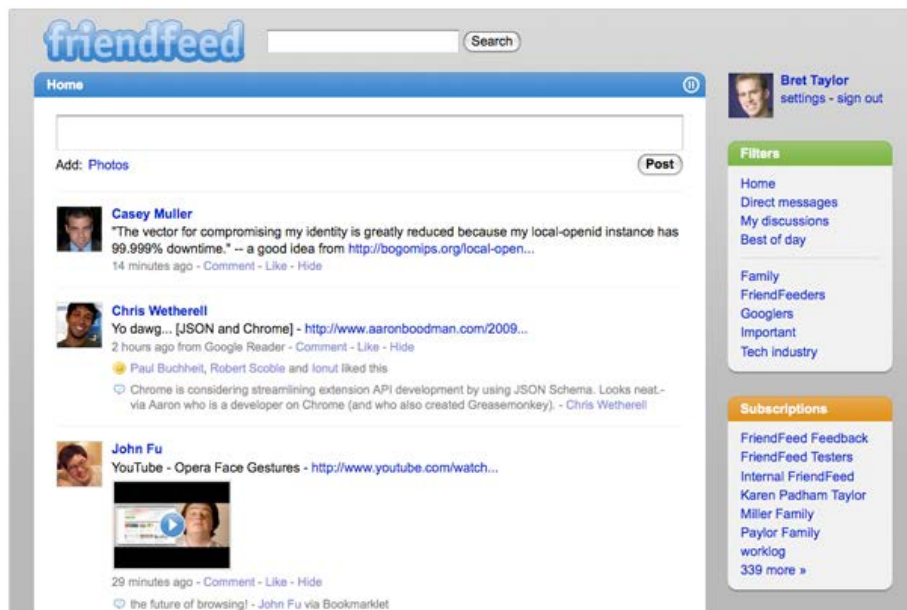
FIGURE 2.2: FriendFeed screenshot

the same social activity. The user performs the social activity within a social network aggregator and the information is synchronized to all of SNS that the user specifies [149].

Currently, both free and commercial social network aggregators are available. All of them reply on APIs provided by the SNS providers for performing the aggregation process. More specially, users have to authenticate their social accounts to be syndicated, and give suitable permissions to the social network aggregator such that it can access to these accounts for collecting data. Once access is granted, recent social data will be regularly or in real-time mode pulled from SNSs into the social network aggregator.

FriendFeed[24] (see Figure 2.2), Hootsuite[25] (see Figure 2.3), TweetDeck[26] (see Figure 2.4) are some representative examples. FriendFeed is used for individual purpose. It aggregates updates, posts and photos submitted by a user and friends on multiple SNSs so that the user can read, share, and comment on these things in real-time without leaving the platform. Hootsuite and TweetDeck are more professional use oriented, and may be used as social network management tools. In addition to the aggregation functionality, they include advanced features like scheduling posts and share in advance, content analysis, bookmark, RSS feeding which allow businesses and organizations to efficiently lead their marketing campaign across SNSs. We will discuss about the drawbacks of the current social network aggregators later in this chapter.

---

[24]FriendFeed : http://friendfeed.com/

[25]Hootsuite : https://hootsuite.com/

[26]TweetDeck : https://tweetdeck.twitter.com/
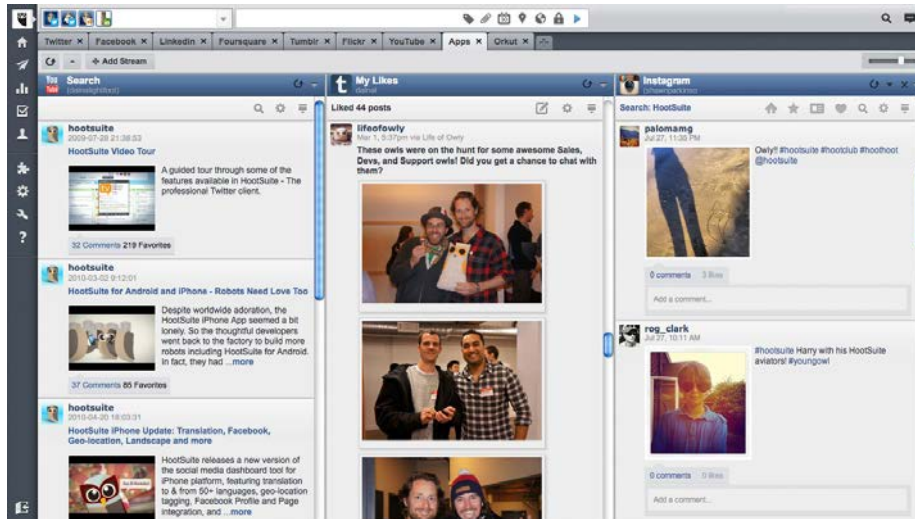
FIGURE 2.3: Hootsuite screenshot
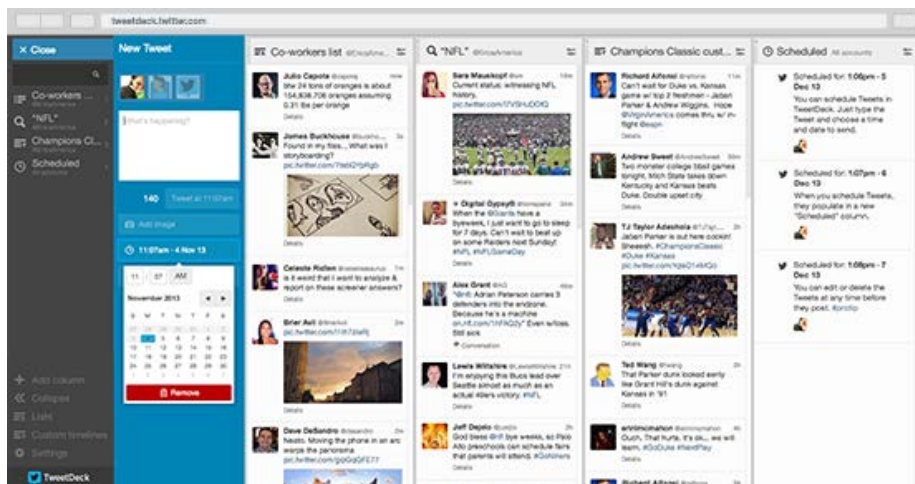
FIGURE 2.4: TweetDeck screenshot

2.2 Information Filtering

Information Filtering (IF) is an efficient solution to the information overload problem. The aim of IF is to expose users only to information that is relevant to them [65]. Many IF systems have been developed for various domain applications, for example personal email, content-sharing platforms, and e-commerce websites. Traditional IF systems typically share the following features [18] :

- They are designed for unstructured or semi-structured data, for example journal articles, email messages;

- They deal primarily with textual data;

- They handle large amount of data;

- They involve streams of incoming data;

- They are based on descriptions of individual or group information preferences;

- They are often meant to imply the removal of data from an incoming stream, rather than finding data in that stream.

Social data meet all these conditions, as they are unstructured, mostly textual, and constantly appearing within a user's social streams in a great number. Thus, many IF solutions adapted to social data have been proposed. We classified them into four major categories: (1) *Friend Grouping*, (2) *Streaming Categorization*, (3) *Stream Filtering*, and (4) *Stream (Re)Ranking*. The (2) and (3) categories are based on the derived description of social data and the user's preferences. The (1) and (4) categories focused on the context of social data (e.g. the provenance, the type of relationships, the common interests, etc.) rather than their descriptions.

2.2.1   Friend Grouping

Friend Grouping approaches consist of splitting a user's entire list of friends into a number of sub-lists of "homogeneous" friends according to certain criteria. So, instead of managing the whole long stream of information coming from all friends, the user can separately monitor different reduced streams via the corresponding sub-lists. The user can check only the contents from a given sub-list at a given time for seeking certain information. The more coherent the lists are, the higher the chances to locate good contents.

Friend Grouping have been already included in some SNSs such as Facebook, Twitter, and Google+. While Facebook and Twitter call it "Lists", Google+ use the term "Circles". Users have to manually create a number of lists and insert into each list different friends according to their personal convenience. These lists are expected to provide user with a primary tool for privacy control, selective sharing and filtering. However, the high burden of manual grouping still prevents many users from adopting it.

Thus, there is a need for automating group creation while allowing users to edit created groups. Facebook includes a feature called "Smart Lists" which are lists automatically generated based on friends' personal information, namely work, school, family and city. For example, a list of friends living in the same town as the user, or a list of friends going to the same school as the user, is automatically generated and updated whenever a new friend is added. However, these lists only cover a small part of the user's entire friend list, and are in some cases too broad to facilitate information filtering.

Gao et al. [57] and Eslami et al. [48] proposed to use graph-based techniques for clustering friends. They first rebuilt for a given user his/her friend graph in which each node is a friend of the user. If two friends are also friends to each other, they are linked. Then, they performed a specific clustering algorithm on the same graph in order to cluster its internal nodes. There are three different categories of clustering algorithms: (1) *disjoint clustering algorithms* where each friend can only belong to one group; (2) *overlapping clustering algorithms* where a friend can be a member of more than one group; (3) *hierarchical clustering algorithms* which categorize friends in a multi-level structure where one group can be a subset of another group.

Qu et al. [120], in addition to using network information (i.e. users' social links), used textual information (i.e. users' tweets published on Twitter) for group member suggestions. More specially, they tried to capture and to model users' topical interests. For that purpose, Latent Dirichlet Allocation (LDA) [23] was applied to derive topics from users' tweets. The proposed system furthermore took group seeds (i.e. some first group members) provided by users so that it could calculate the similarity between one target user and group seeds in order to determine how likely the target user belong to the group in question.

In the real world, users like to group their friends in many ways [120]. For example, some might create topical lists like "computer scientists" or "television comedians" while others might create lists containing their real-life friends. So, it is probable that automatically generated lists are not enough personalized to meet any user's requirements. Additionally, as one's interests change with time [121], users still have to maintain and adjust regularly their lists.

2.2.2 Stream Categorization

Text Categorization, also known as text classification, or topic spotting, aims to automatically sort a set of documents into categories (or classes, or topics) from a predefined set [132]. Each document can be either in multiple, or exactly one, or no category at all [77]. Users can subscribe to and view only the documents of certain categories based on their own interests. *Stream Categorization* applies Text Categorization to a user's social stream where each piece of social data is considered as a document.

The first attempt of Stream Categorization was introduced by Sankaranarayanan et al. in [129]. Their system, called TwitterStand, is intended to classify incoming tweets as either *junk* or *news* where the junk tweets have a good chance of not being related to the news and hence, are discarded, while the news tweets have a good chance of being related to news. For that purpose, they used a naive Bayes classifier [92] that was trained on a training corpus of tweets that had already been marked as either news or junk. A very similar
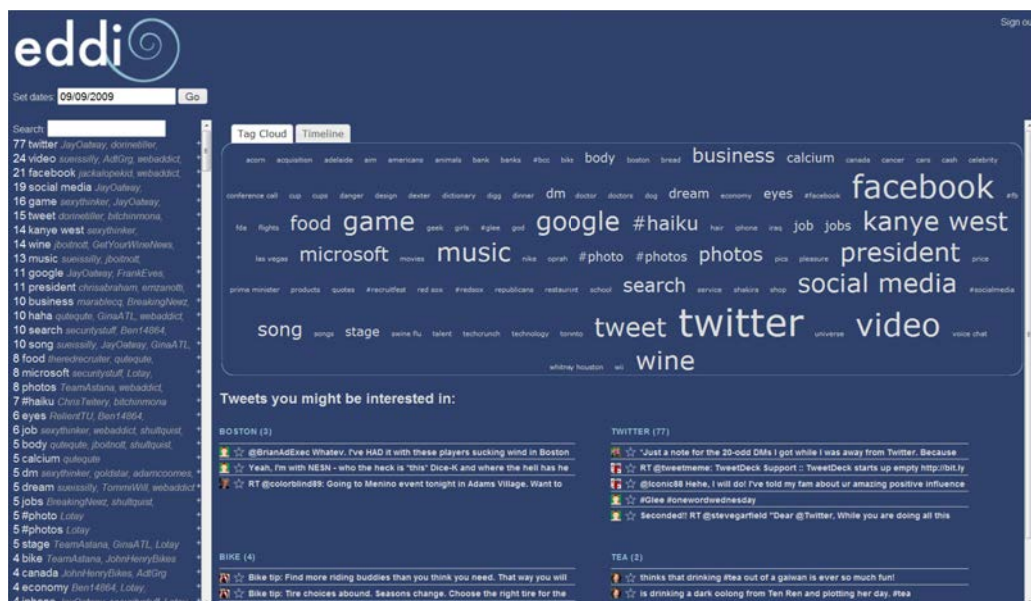
FIGURE 2.5: Eddi - Interactive Topic-based Browsing [21]

but more general work was introduced in [143], where the authors increased the number of categories from 2 to 5 categories such as *news*, *events*, *opinions*, *deals*, and *private messages*.

Categorizing social data by predefined topics may encounter two particular issues, as social data are various, numerous and constantly evolving. First, if the topics are too broad, the associated contents will overfeed the user. Second, if the topics are conversely too specific, they will become outdated quickly.

To address such dynamic, broad nature of social data, some researchers [4, 21] proposed to replace predefined topics by dynamic topics. More specially, they applied *Topic Identification* techniques to identify the topics of each piece of incoming social data. Moreover, taking into consideration the fact that social data are often short and ungrammatical texts, researchers have also used external sources of information to enrich and disambiguate the original content of social data.

For example, Bernstein et al. [21] proposed to use search engines to identify the topics of a tweet. A tweet is first transformed into keywords, mostly noun phrases, which are then sent as a query to a given search engine in order to retrieve a result set of documents (i.e. top ten results). From those results, a number of most frequent noun phrases, including those requested, are extracted and considered as the topics of the tweet.

Alternatively, Abel et al. [4] used DBpedia as an external knowledge source to enrich users' tweets. Their technique consists of extracting named entities from the tweets. The identified entities are linked to the corresponding DBpedia entities, the categories of which (e.g. locations, people, events, etc.) are also retrieved. Both the DBpedia entities and their categories are assigned to the original tweets as topics.

At visualization level, most of researchers have adopted a faceted browsing interface where each facet corresponds to a given topical category, for instance Eddi (see Figure 2.5).

2.2.3 Stream Filtering

Stream Filtering aims at actively delivering users with only information relevant to their interests. Unlike the previous method, Stream Categorization, where a user explicitly specify his/her interests by selecting certain categories, Stream Filtering attempts to implicitly learn and infer the user's interest profile from the contents that he/she published in SNSs. In general, incoming information is selected by how well it, or more precisely its description, matches the user's interest profile. Technically, the user interest profiles and the information descriptions are represented by a set of weighted items (e.g. keywords, hashtags, topics, semantic entities) [2].

Chen et al. [38] used the traditional Bag of Words and TF-IDF weighting to generate and weight users' interest keywords from their tweets. They modelled incoming tweets, in particular those containing URL, by the words inside the tweets and then used cosine similarity to decide whether a given tweet was in the scope of the user's interests or not. The main problems with this method are Polysemy, which is the presence of multiple meanings for one word, and Synonymy, which indicate that relevant information can be missed unless the exact keyword exist in the profile [2].

Garcia-Esparza et al. [59] proposed another user profiling approach based on the topical categorisation of users' posted tweets. Like [38], they mainly focused on tweets containing URL. These tweets are assigned to one or more categories based on the contents of the referred webpages using a naive Bayes Multinomial classifier [83]. There were 18 categories in total, which correspond to general topics such as music, sports or health. A user's interest profile is derived from the categories of his/her posted URLs, and is used as the basis for filtering his/her timeline, prioritising those tweets that conformed to the user's own interests (as illustrated in Figure 2.6).

Kapanipathi et al. [80] applied a semantic approach to construct users' profiles. Using the same technique described in [105], they extracted entities from users' tweets and linked them to DBpedia concepts. Each extracted concept represented a user interest of which the weight was calculated using the frequency of occurrences. Kapanipathi et al. also collected interests that had been explicitly stated by the users on LinkedIn and Facebook.
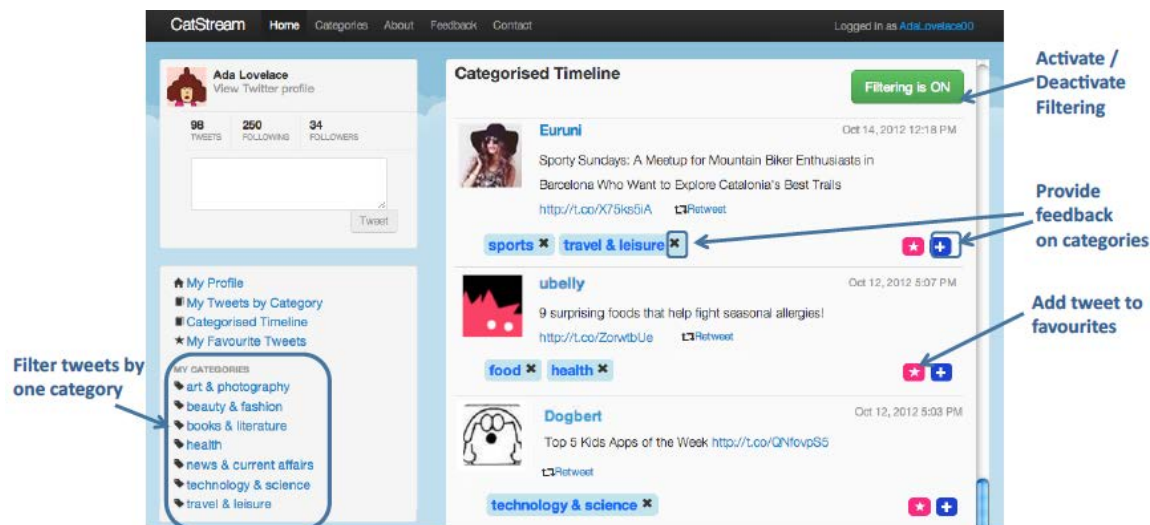
FIGURE 2.6: CatStream categorized timeline [59]

2.2.4   Stream (Re)Ranking

Most of SNSs display a user's social stream in reverse chronological order. Such a stream is full of contents of very different qualities, many informative or relevant contents might be flooded or displayed at the bottom while some nonsense buzzes might be ranked higher [136]. *(Re)Ranking* approaches consist of (re)ordering the user's social stream in prioritizing the information relevancy.

Facebook implemented its own ranking algorithms, called EdgeRank. This algorithm decides which stories called "edges" (e.g. status updates, comments on another status update, photo tags, etc.) appear in each user's homepage. EdgeRank is proprietary and not available to the public. No one other than Facebook knows exactly how it works. Nevertheless, Facebook revealed in 2010 the three following ingredients of the algorithm [151] :

- *Affinity Score* means how "connected" a given user is to the Edge, or more precisely its author.

- *Edge Weight* shows that each category of edges has a different default weight, for example comments are heavier than likes.

- *Time Decay* indicates that a story loses points, as it gets older.

The final rank of an edge is therefore computed from these three scores. Facebook will filter each user's homepage to only show the top-ranked stories for that particular user. Seemingly, the stories of friends who interact the most with the user often go to the top of the stream.

Researchers also started to be interested in social stream (re)ranking. Several re-ranking methods have been introduced for different SNSs, for example Twitter [136], LinkedIn [11, 71], and Facebook [28]. Like EdgeRank, these methods do not focus on the analysis of the textual content of incoming information, but rely on a wide set of specific features involving not only the information, but also the source and the target users.

Shen et al. [136] included the freshness of tweets, the influence of authors, the quality of tweets, and some social features. The freshness of tweets is the difference between the time t the user u saw the tweet m and the time this tweet was posted or the rank of the tweet m in the user's timeline at the time t when the user visited Twitter. The influence of authors is computed based on the number of followers that the author a has, the number of users that the author a follows, the number of lists the author a belongs to, the number of tweets the author a posts each day, the number of days since a's account was created, and whether the author a is authenticated by Twitter officially or not. The quality of tweets is determined by taking into account the length of the tweet m, whether the tweet m contains URL, the number of hashtags that appear in the tweet m, the number of retweets that rooted from the tweet m. Social features represent the relationship between the user u and the author a which is computed with the number of a's tweets being retweeted by u in the past, the number of a's tweets being replied by u in the past, the percentage of a's tweets retweeted by u, the percentage of a's tweets replied by u.

Bourke et al. in [28] introduced some additional interesting features for re-ranking a message posted on Facebook. These are, for instance, the number of explicit clicks the message received, the number of shares the message received, the number of comments the message received, and the mean number of message comments per hour and the number of comments received in the last hour.

Generally, based on a set of predefined features, a ranking model is first built on a training dataset with a certain machine learning technique, and is then applied to incoming information to compute its corresponding rank.

2.2.5 Summary

Table 2.2 summarizes the principal techniques, on which the related works of the four approach categories (i.e. *Friend Grouping*, *Stream Categorization*, *Stream Filtering* and *Stream (Re)Ranking*) rely, and their respective limitations.

TABLE 2.2: **Summary of Information filtering related work**

Approach	Technique(s)	Related work	Limitation(s)
Friend Grouping	Manual grouping	Facebook lists, Google+ circles, [160]	Regular maintenance required
	Graph-based clustering algorithms	[48, 57, 120]	Not personalized enough for users' various preferences
Stream Categorization	Text categorization with the predefined topics	[129, 143]	Model training required, Too generic topics
	Topic detection	[4, 21]	Too many topics generated
Stream Filtering	User profiling based on users' social activities	[38, 59, 80]	Incomplete and/or imprecise profiles for passive users
Stream (Re)Ranking	Ranking algorithms based on a set of specific features	[28, 136, 151, 160]	Model training required, Sources containing helpful contents but low-ranked

2.3 Information Discovering

Information Discovering is loosely used here to categorize various solutions. Briefly, the solutions of this categorization are expected to help users discover new and interesting information from outside of their social streams. We arranged related works into two main groups : (1) *Friend Recommendation* and (2) *Additional Sources*.

2.3.1 Friend Recommendation

The aim of *Friend Recommendation* is to recommend to a user other members who are likely interesting to follow. Most of current SNSs include a friend recommendation feature. For example, Facebook, LinkedIn and Twitter all show in every user's homepage a suggestion panel titled "people you may know". However, they are roughly based on the number of common connections that two different users have, in order to suggest one to another.

Hannon et al. [66] introduced a user profile based recommender for Twitter users, called Twittomender. The proposed system attempted to build up for each user an interest profile which is a weighted term-vector. The frequent terms are directly extracted from the tweets published by the user and/or friends. Upon one's request, his/her profile would be matched with others' profiles in order to find out the most similar ones.

Likewise, Armentano et al. [15] proposed an algorithm for recommending followees in Twitter. In addition to using users' profiles inferred from their tweets, they moreover added an extra selection step at the begining of the recommendation process. In fact, the

proposed algorithm limited the candidates to those inside one's extended social network (i.e. friends of friends). Their main assumption was that if a user u_F follows a user that is also followed by u_T , then other people followed by u_F can be of interest to u_T.

We may also assign the works of Weng et al. [150] and Lim and Datta [94] to the field of Friend Recommendation. Weng et al. [150] proposed a new algorithm called TwitterRank to measure the influence of Twitter users. TwitterRank took the link structure between users, and especially users' expertise and/or interests into account. It was therefore intended to find out topic-sensitive influential users. Lim and Datta [94] differed with an approach for finding communities of users with common interests. They first identified celebrities that were representative of an interest category, then detect communities based on linkages among followers of these celebrities. Although, these works do not directly suggest suitable friends to a user, they provide the user with open choices to consider based on his/her own interests.

Friend Recommendation is useful for a user who wants to discover new and interesting people to follow, thus reaching relevant information. It works and is effective when the user is new to the SNS or has few friends. Otherwise, it becomes counter-productive, as the user has to manage with many friends, and all the contents that those friends may share.

2.3.2 Content Recommendation

Content Recommendation approach is complementary to *Stream Filtering*. Stream Filtering solutions limit the user experience only to his/her personal stream, which means that the information sources are restricted to the user's circle of friends. Additional Sources then allows a user to increase the number of sources without expanding his/her current circle of friends which is already very large.

Chen et al. [38] extended a Twitter user's sources of information in two ways *FoF* (followees-of-followees) and *Popular*. FoF takes the tweets of the users who are followed by the users who are followed by the target user. Popular takes popular URLs which are the most shared by Twitter users.

Kapanipathi et al. [80] built a central repertory called "Social Hub" which allows users to subscribe and regularly receive interesting tweets which are extracted from the entire Twitter public stream.

Other researchers [6, 111] recommend a user the latest news articles (e.g. posted within 24 hours) from The Huffington Post [111], from BBC, CNN or New York Times [6] which are related to the user's news interests derived from his/her tweets.

2.3.3 Summary

Table 2.2 summarizes the recommendation strategy, especially the choice of sources, of the *Information discovering* related works, and their probable limitations.

TABLE 2.3: **Summary of Information discovering related work**

Approach	Source(s)	Related work	Limitation(s)
Friend Recommendation	Members of the same SNS	[15, 66, 94, 150]	Information overload risk
Content Recommendation	Contents published by friends of friends	[38]	Incomplete and/or imprecise profiles of inactive users
	Contents published by the entire SNS	[80]	Incomplete and/or imprecise profiles of inactive users
	Contents form external sources	[6, 111]	Incomplete and/or imprecise profiles of inactive users

2.4 Discussion

Thus, we have seen a number of works related to Social Network Aggregation, Information Filtering, or Information Discovering (see summary table 2.4). These interesting works have their preferred way to help users to better manage their social networks and the information shared within them. Nevertheless, they also have some drawbacks.

Current social network aggregators allow to 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. Some of them are able to display contents into different categories based on their types, for example photos, videos, links, updates, and so on. It unfortunately does not help users very much to extract useful information.

Information Filtering is important as its allows to reduce users' filtering efforts and provide interesting contents. However, most of related works have applied different machine learning techniques, which need to be trained on some set of training data. Consequently, they have become domain-specific solutions. For example, an efficient solution dedicated to Twitter is no longer suitable for Facebook. All presented related work, discussed above except [160], have actually been designed for only Twitter or only Facebook.

Information Discovering is complementary to Information Filtering. It is interesting in that it helps a user to discover additional information from outside his/her circle of friends. Some related works, presented above, have considered additional sources of information only from a given SNS, or from very popular news media. Groups of interests have not been considered yet, even though they represent a very reliable source of contents of interest.

One of the most interesting works is [160] where the authors introduced a personalized social network aggregator and recommender, named SocConnect, supporting at this moment the two SNSs Facebook and Twitter. In order to filter the heterogeneous social data from both SNSs, the authors have implemented manual Friend Grouping and applied different machine learning techniques on the textual features and the non-textual features (e.g. actor, activity type, source, etc.) of social data as well. SocConnect is however limited to an individual basis. The users are not expected to share the interesting information and useful contents with each other.

This is also the case for all the above-discussed works. None of them included features in response to the *Sharing* question, as shown in Table 2.4. Certain existing collaborative systems may include features allowing to retrieve and capitalize public contents from social networks, essentially Twitter, by watching specific keywords. However, they need to permanently listen to the entire social network, which furthermore leads to a big number of contents to review and to filter manually or automatically. The most suitable answer is at the moment to encourage the members of a given group to manually select and put the useful contents from their respective social streams into the group so that other members can access to as well.

No complete answer to the three questions of filtering, discovering, and sharing social data across SNSs has been proposed yet. Our work therefore aims at searching for such an answer. In the following chapters, we will see in detail how our answer is conceptually designed (Chapter 3 - *Conceptual Design*) and technically implemented (Chapter 4 - *Technical Solution*).

TABLE 2.4: **Summary of related work**

Reference	Support multiple SNSs	In response to Q_1-Filtering question	In response to Q_2-Discovering question	In response to Q_3-Sharing question
FriendFeed, Hootsuite, TweetDeck, Plumbaum et al. [118]	yes	no	no	no
Gao et al. [55], Eslami et al. [48]	no: only Facebook	yes: based on Friend Grouping	no	no
Qu and Liu [120], Rakesh et al. [121]	no: only Twitter	yes: based on Friend Grouping	no	no
Sankaranarayanan et al. [129], Sriram et al. [143]	no: only Twitter	yes: based on Stream Categorization (static categories)	no	no
Bernstein et al. [21], Abel et al. [4]	no: only Twitter	yes: based on Stream Categorization (dynamic categories)	no	no
Shen et al. [136]	no: only Twitter	yes: based on Streaming (Re)Ranking	no	no
Bourke et al. [28]	no: only Facebook	yes: based on Streaming (Re)Ranking	no	no
Hannon et al. [66], Weng et al. [150], Lim et al. [94], Armentano et al. [15]	no: only Twitter	no	yes: based on Friend Recommendation	no
Chen et al. [38]	no: only Twitter	yes: based on Streaming Filtering (keyword level)	yes: URLs from followees of followees and popular URLs on Twitter	no
Garcia-Esparza et al. [59]	no: only Twitter	yes: based on Streaming Filtering (topic level)	no	no
Kapanipathi et al. [80]	no: only Twitter	yes: based on Streaming Filtering (semantic entity level)	yes: all public tweets	no

Reference	Support multiple SNSs	In response to Q_1-Filtering question	In response to Q_2-Discovering question	In response to Q_3-Sharing question
Abel et al. [6], O'Banion et al. [111]	no: only Twitter	no	yes: news articles published by news media	no
Zhang et al. [160]	yes: Twitter, Facebook	yes : combines Friend Grouping (manual) and Stream Filtering and Stream (Re)Ranking	no	no

Chapter 3

Conceptual Design

Contents

In the previous chapters, we have raised the three following questions:

1. Q_1-Filtering: *How to help users to extract contents of interest from their different social networks with less effort and without altering their social networking experience?*

2. Q_2-Discovering: *How to help the users to discover additional contents of interest from outside of their cycles of friends?*

3. Q_3-Sharing: *How to make it possible for the users to share the contents of their various social streams with their respective groups without extra charge?*

To answer these questions, we propose in this work a novel ***User-centered and group-based approach for social data filtering and sharing***. This approach is in line with 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 [106]. A SIS enables strategic applications whose main strength is the integration of different communities that nevertheless preserves their diversity and autonomy [33].

Conceptually, our approach consists of two main components: (1) *User-centered social data filtering*, and (2) *Group-based social data sharing*. The first component is meant to answer the question Q_1 by allowing the users to aggregate their different social networks and to extract and organize contents which are of the users' interests. The second component is intended to answer both questions Q_2 and Q_3 by enabling collaborative spaces over SNSs where the members of a given group of interest can share with each other the contents of interest originated from their respective social networks.

In this chapter, we will first present the general requirements that these two components must fulfill. Then, we will go into the details of their respective conceptual foundations. To conclude this chapter, we will stress out how our approach meets the expectations.

For reasons of clarity and consistency, from here we will use the term "social data" to refer to a user's raw social data aggregated from different SNSs and the term "contents of interest" or simply "contents" to refer to the social data which are processed and considered to contain information relevant to the user's interests.

3.1 General Requirements

3.1.1 Filtering Requirements

The question Q_1 leads to the three underlying questions:

- $Q_{1.1}$: *How to access and collect a given user's social streams, which are protected by and scattered across different SNSs?*

- $Q_{1.2}$: *What are the appropriate techniques for extracting the contents, which are relevant to the user's interests, given that the available contents are numerous and mostly text-based?*

- $Q_{1.3}$: *How to organize the contents of interest and how to present them to the user?*

A solution noted S supposed to answer the question Q_1 and taking into consideration its three underlying issues, could be in compliance with the overall design illustrated in Figure 3.1. S is at least composed of two features (1) *Aggregating* and (2) *Filtering*. The first feature answering the question $Q_{1.1}$, is responsible for aggregating a user's social data across SNSs in a single place. It should also comply with the privacy policy of SNSs.
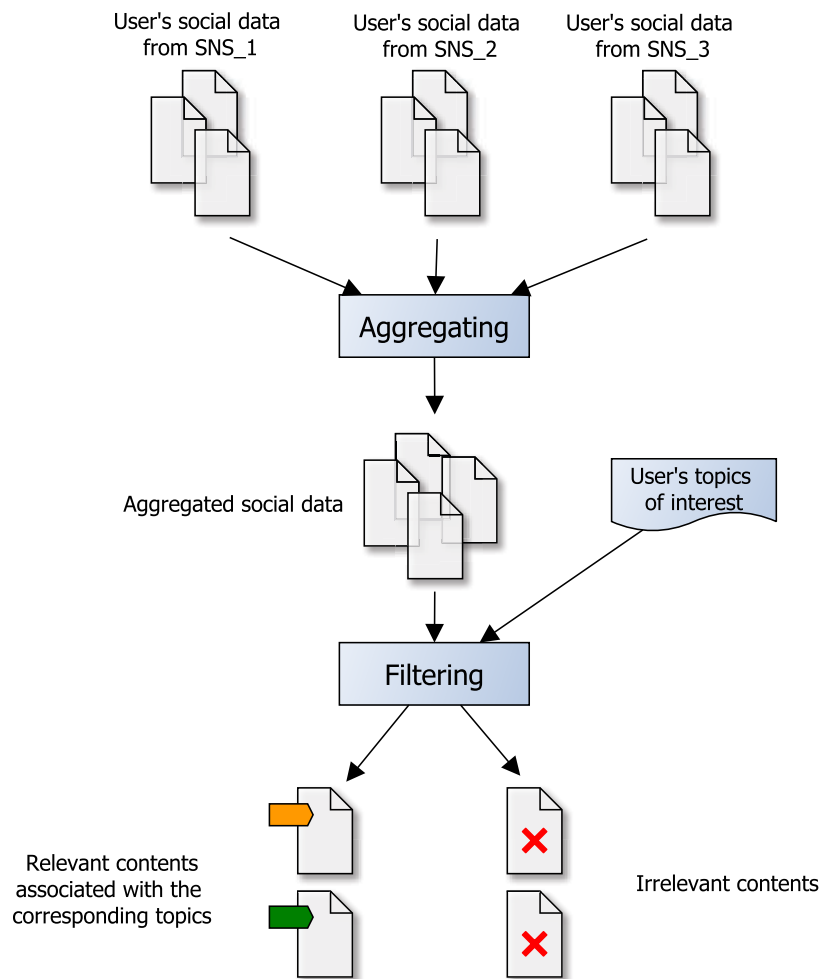
FIGURE 3.1: User-centered social data aggregating and filtering overall design

The second feature addressing the questions $Q_{1.2}$, $Q_{1.3}$ is responsible for helping the user to extract the contents of interest from the aggregated social data with a minimum of manual effort. More specially, each piece of social data is analysed and classified as interesting contents or useless data. Only interesting contents are selected for being shown to the user, whereas others will be hidden from the user's view. The contents of interest should furthermore be organized by topics which means that each of them is indexed and associated with one or several topics of interest. This way, the user can easily and quickly access to expected information by selecting the corresponding topic.

Most importantly, it is required that both aggregating and filtering features of S have a user centric approach. A user-centered aggregating feature should straightforwardly ask the user for authenticating and authorizing access to his/her social data across SNSs, thus being able to recover an extended range of social data. A user-centered filtering feature should be semi-automated, which means that the topics of interest are, instead of implicitly
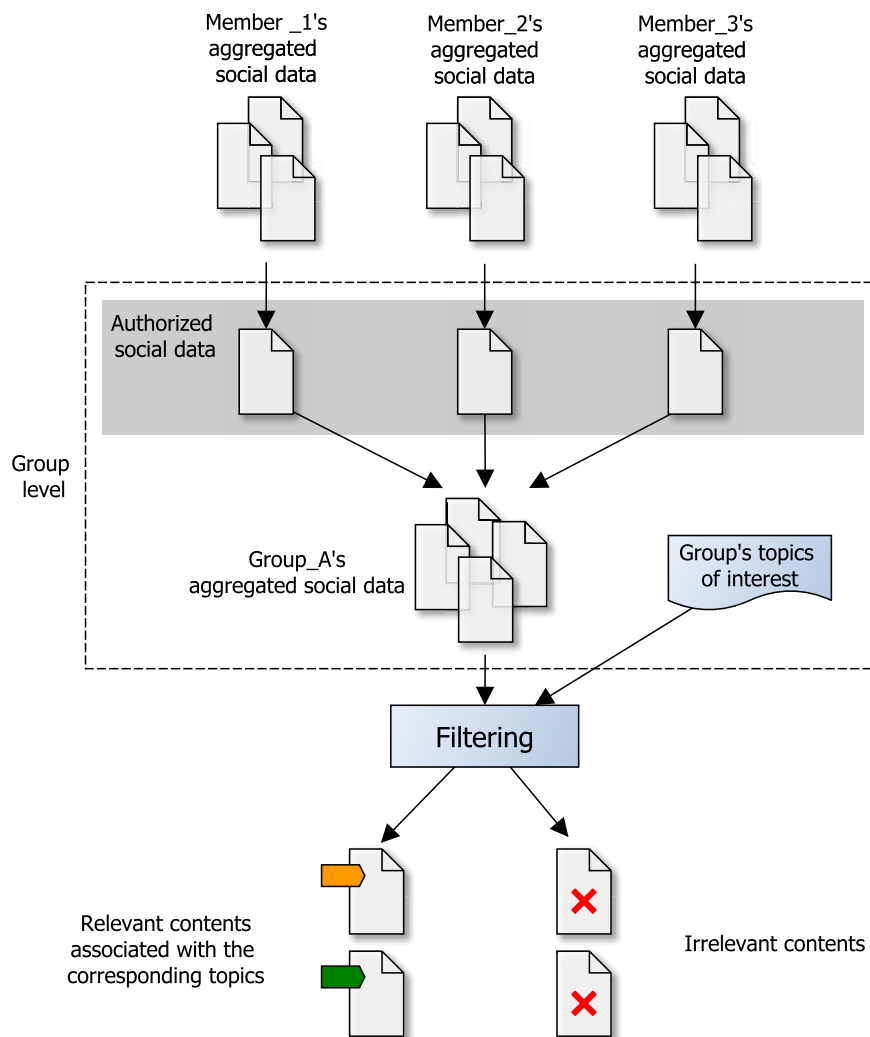
FIGURE 3.2: Group-based social data discovering and sharing overall design

learned, explicitly and gradually provided by the user while the social data are automatically processed and filtered according to these topics.

3.1.2   Sharing Requirements

The two questions Q_2-Discovering and Q_3-Sharing can actually be solved by a common solution noted S^+, which is an extension of the solution S as illustrated by Figure 3.2. The most important requirement for extending S to S^+ is to embed in S an extra organization level enabling collaborative spaces devoted to groups of interest. Such a group setting allows to achieve the two objectives, *discovering* and *sharing*, without many changes (see Figure 3.2).

The users are able to join a given group and contribute to build up its collective social data sources by sharing their own social data previously retrieved by the aggregating feature of S with the group. Then, the same semi-automated filtering feature of S would be applied to the group's collective social data sources to automatically discover the contents of interest. The only difference is that a group's topics of interest are not defined by one person, but collectively defined by any interested member of the group in order to take advantage of the expertise of everyone. This way, taking part in a group of interest allows a user to discover additional useful contents (Q_2), and to share with his/her respective groups of interest (Q_3).

Like S, the solution S^+ must also respect the user-centered requirement. More specifically, a member of a given group should be able to choose which part of his/her social data can be shared with the group (see Figure 3.2), as he/she may not want to unveil some sensible information to the group.

3.2 User-centered Social Data Filtering

In this section, we will present two conceptual foundations of the *user-centered social data filtering* component. First, we will introduce a model serving as a base for integrating the heterogeneous social data from different SNSs. Then, we will describe the principles of filtering and organizing contents of interest.

3.2.1 Social Data Integration

3.2.1.1 Social Data Scope

Until now, we have used the term "social data" in a quite ambiguous manner to indicate data generated by users as well as data related to a particular user in SNSs. From here, we will only refer to social data as belonging to a particular user and defined as:

> *"A user's social data include contents published by, or involving, or shared with the user in the social network sites to which he/she is connected."*

This definition comprises the information that a user pushes on the SNSs and the information that he/she receives from his/her social friends. It corresponds to the sum of the user's social streams in their basic version (i.e. before being filtered and/or ranked), and includes a wide range of information such as profile information, social connections, postings, interests, and so forth. The social data vary from one SNS to another. For example, the user

TABLE 3.1: Social data available via the APIs provided by different SNSs

Attribute	Facebook	Twitter	LinkedIn	Google+
nickname	x	x	x	x
first name	x		x	x
last name	x		x	x
full name	x	x	x	x
profile photo	x	x	x	x
about	x	x	x	x
email	x		x	x
homepage	x	x	x	x
location	x	x	x	x
gender	x			x
birthday	x		x	x
relationship status	x			x
language	x	x	x	x
education	x		x	
affiliations	x		x	x
interests	x		x	
groups	x		x	
publications			x	
project			x	
contact	x		x	
social connections	x	x	x	x
posts	x	x	x	x

profile on Twitter is currently very limited. It only includes name, bio and location of the member. The user profile on Facebook is more elaborate. 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 quotations;

- Contact information such as mobile phone, landline phone, school mailbox, address, etc.;

- Education and work information such as the names of schools attending/attended, and current employer.

Therefore, we had to determine a suitable scope of social data for our study. We closely investigated the top SNSs, namely Facebook, Twitter, LinkedIn, and Google+, for social data made available via the provided APIs (see Table 3.1). From this study, we have identified the following six most frequent and important information dimensions:

1. The *Profile Information* dimension includes basic information about the user such as name, about, language, email, gender, location, etc.;

2. The *Friend* dimension represents the social connections established between the user and other members of a SNS;

3. The *Group* dimension lists the groups, created on a SNS, that the user is a member of;

4. The *Studie & Work* dimension describes respectively the school and academic experience and the professional experience of the user;

5. The *Interest* dimension lists the user's interests, often explicitly claimed by the user;

6. The *Post* dimension represents the contents published by as well as those shared with the user.

Note that such dimensions are not completely exclusive to each other and that there may be some overlaps between them. For example, people often join a specialized group because they share some common interests with the group, or a group can be formed by people from the same school or the same workplace.

3.2.1.2 Integration Model

Each SNS utilizes its own syntax and terms for representing social data. It is therefore very common that different terms are used for the same type of data. For example, a piece of text published by a user is called "tweet" on Twitter but "post" on Facebook, or a social contact is called "friend" on Facebook but "connection" on LinkedIn. Given such diversity, a common model is necessary for integrating the heterogeneous social data from different SNSs. Furthermore, in order to cover all the aforementioned dimensions (i.e. *Friend*, *Group*, *Study & Work*, *Interest*, and *Post*), we have built an adapted integration model based on FOAF [32] and ActivityStream [139] as illustrated by the UML class diagram in Figure 3.3.

A *user* is a person ($\sim$ `foaf:person`) who is identified by his/her unique email address. The user can have several *social accounts*, each of which is from a different *social network*, for example a user can hold a Facebook account, a LinkedIn account and a Twitter account. The association between *User* and *SocialAccount* is equivalent to the semantic property `owl:sameAs`.

Each social account contains a number of attributes identical to the relatively invariant information of the two dimensions *Profile Information* and *Studies & Works* (see Figure 3.8). Other types of social data (i.e. *Friend*, *Group*, *Interest*, and *Post*), which are changing
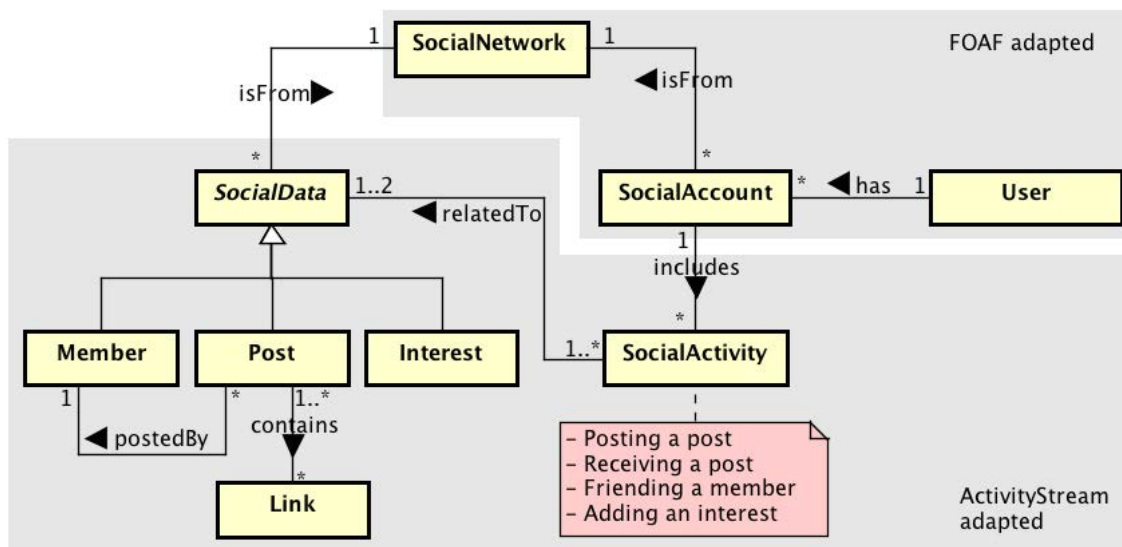
FIGURE 3.3: Social Data Integration Model

over time, are linked to the social account through a number of timestamped *social activities* taken in the same social network as the social account.

ActivityStream defines many possible social activities. Nevertheless, all of them are not necessary and include useful information. At this time, we have only needed four specific types of *Social Activity* namely "post", "receive", "befriend", and "add".

Each type of social activity refers to one or even two given subclasses of the abstract class *Social Data*:

- The *post* activities are related to the *Post*-type social data, for example, a user via his/her social account *posts* a *post* on a given social network,

- The *receive* activities are related to the *Post*-type and the *Member*-type social data, for example, a user via his/her social account *receives* a *post* which is posted by another *member* of the same social network,

- The *befriend* activities are related to the *Member*-type social data, for example, a user via his/her social account *befriends* with another *member* of the same social network,

- The *add* activities are related to the *Interest*-type social data, for example, a user via his/her social account *adds* a new *interest*.

The *Post*-type social data corresponds to the *Post* dimension. The *Member*-type social data has a larger scope than that of the *Friend* dimension. It includes normal members whom a social account can be friend of, and special members whom a social account can

only follow to receive their posts, for example pages on Facebook, or company accounts on LinkedIn. The *Interest*-type social data incorporates both *Interest* and *Group* dimensions.

The social activities are unique to the corresponding social account, whereas the social data are unique to the original social network. This means that some social activities from different social accounts may refer to a same piece of social data. For example, two social accounts befriend with a same member, thus receiving the same posts from this member.

Every subclass of *Social Data* contains at least one text-valued attribute. The **Member** class includes a description attribute. The *Interest* class includes a name and a description attributes. The *Post* class contains a text attribute. In the case where a post contains one or several links, the title and description of the referred webpages are considered as the extended text-valued attributes of the post as well. The text-valued attributes of each subclass are important, as they provide the description, based on which the social data would be either selected or filtered out with respect to the users' interests during the filtering process.

This generic model can easily be extended. If we later identify some important types of social data and would like to include them into the model, all we will have to do is to add them as subclasses of the class *Social Data* and declare their corresponding *social activities*. Importantly, there will not be any effect on the current model.

For the sake of simplicity, in the rest of this thesis, we will use the term social data in order to refer to the associations of *Social Activity* and *Social Data*, which are actually things that can be filtered and shared. There are therefore four types of associations as shown in Table 3.2.

TABLE 3.2: Associations of Social Activities and Social Data

Association	Social Activity	Social Data
Friend	"befriend"	Member
Post	"post"	Post
Following Post	"receive"	Post $\oplus$ Member
Interest	"add"	Interest

3.2.2 Information Filtering & Organization

The information filtering process consists of constantly analysing any new social data and accordingly taking appropriate decisions to either ignore or show it to a particular user. An information filtering solution should therefore take into consideration the user's *information needs* which reflect his/her short, medium or long-term interests [61]. The user's interests can be gathered in an implicit manner where they are derived from the user's
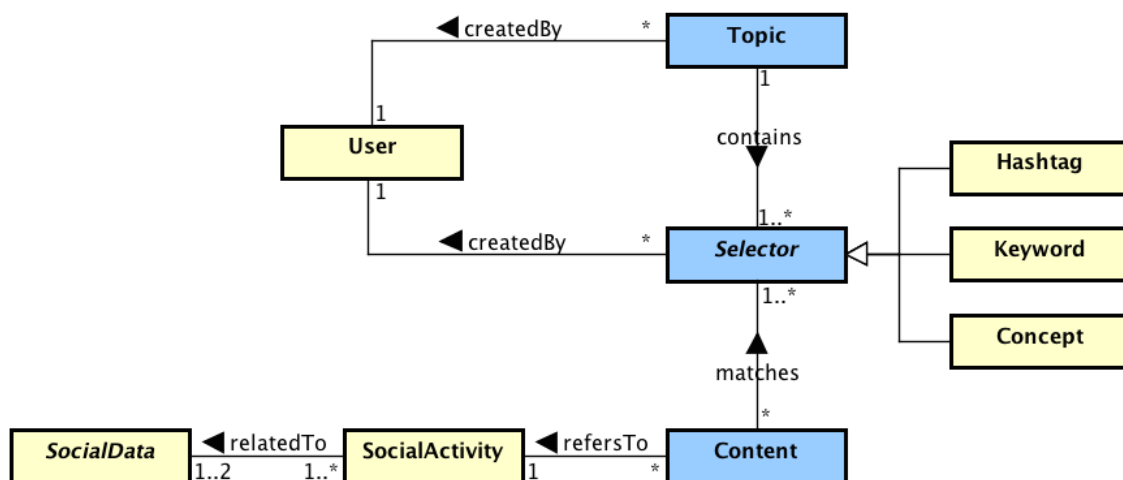
FIGURE 3.4: Topic and selector structure

usage behaviour and history without any extra effort, or in an explicit manner where the user has to explicitly supply them to the system [60].

Unlike traditional filtering systems, in particular recommender systems where data are domain-specific, social data are much more diverse. Java and al. in [76] showed some of the main user intentions while using SNSs:

- Daily chatting about daily routine or what people are currently doing;

- Conversations to comment or reply to their friends' posts;

- Information/URLs sharing;

- News reporting to report latest news or comment about current events.

The generated data thus contain not only helpful information but also useless junk [129].

On the other hand, the interests of users do not follow a simple predictable model. They have a wide range of interests across a large set of topics, even within a topic [101]. Besides, they are often influenced by their social connections, and adapt their own interests in accordance with others' interests. Many of them are *information seekers* who post rarely, but follow other users regularly [76]. 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, we have adopted the explicit way in which the user explicitly provides his/her interests in terms of topics. The user can moreover edit (i.e. add or delete) his/her topics of interests over time. This way, the target system knows exactly what the user needs, thus extracts only the corresponding contents.

To enable the extraction and the organization of contents of interest, we have applied a two-level structure (see Figure 3.4 or Figure 3.8 for more details). The first level called *topics* corresponds to the topics of interest. This level is used to classify and organize the contents of interest. The second level called *selectors* represents the technical specifications of *topics*, which specifies how a piece of content matching a topic should be automatically extracted from the social data. As shown in Figure 3.4, the class *Selector* is expressly left abstract, and could be specified by various information extraction methods. We have proposed at the moment three types of selectors, namely *Hashtag*, *Keyword* and *Concept*. Each type corresponds to a different technique and usage that we will detail further in the subsection *Developed selectors* of the next chapter. A topic can be specified by as many selectors as possible (e.g. two Keyword-based selectors and a Hashtag-based selector) to increase the chance to retrieve helpful contents.

Note that a piece of social data is considered as a content of interest and associated with a given topic when it matches at least one of the topic's selectors. A content can be assigned to several topics. A user can set up (i.e. instantiate with values) as many selectors as he/she desires to a topic in order to increase the probability for detecting interesting information.

3.3 Group-based Social Data Sharing

In this section, we explain how we are conceptually able to extend S to S^+ for supporting *group-based social data sharing* as well.

3.3.1 Group Settings

As mentioned in the *sharing requirements*, to enable the information sharing and discovering, it is necessary to introduce collaborative spaces. We have then added an extra level of organization called *groups*. Furthermore, we have maintained the topic-selector structure, while putting it under *groups* as illustrated by Figure 3.5 (see Figure 3.8 for more details).

In addition, we have specified two kinds of groups:

1. Private groups which are groups only accessible by their creators,

2. Open groups which are groups open to any user for joining.

Such a distinction is interesting, since it enables the user experience on an individual basis, and in a group setting as well. Private groups are mainly dedicated to personalized information filtering, whereas open groups are used for the sharing and discovering purposes. A user can be a member of several whether private or open groups.
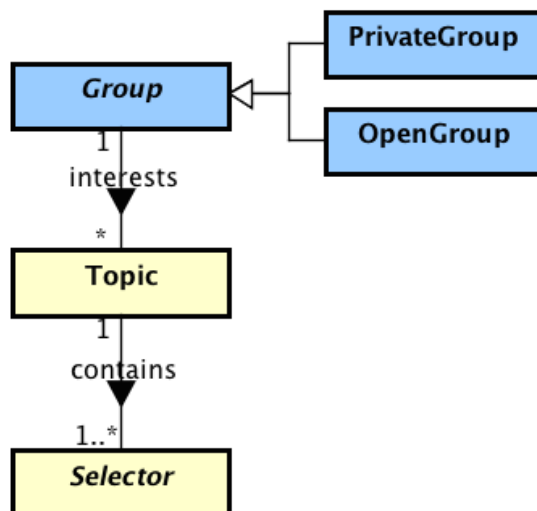
FIGURE 3.5: Group Organization Level
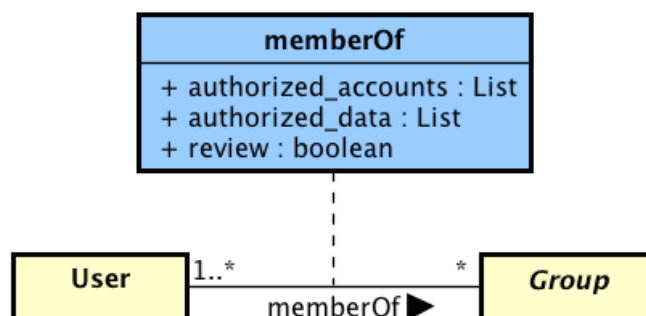
FIGURE 3.6: Sharing settings

In the case of an open group, all members are equal, as there is no particular need for specifying further their different roles yet. However, the fact that one's social data are shared with and visible to others may raise the privacy problem, as some parts of social data may contain sensible information that the user does not want to reveal. It is therefore important to give the user a control over what he/she is ready to share with an open group instead of systematically sharing all of his/her aggregated social data. We have then included features for the users to personalize their membership dues towards each of their groups, as shown in Figure 3.6. The two classes *User* and *Group* are linked through the association class called *memberOf*, which contains following three specific attributes reflecting a member's sharing settings as follows:

1. *Authorized accounts* indicates which social accounts and their associated social data can be matched with the group's topics of interests and eventually shared with the group members;

2. *Authorized data* indicates which types of social data can be used for sharing;

3. *Review* if enabled, prevents a newly detected relevant content to be immediately shared with the group but waiting for the user's approval.

The first and second attributes allow a user to restrict the sharing scope of the social data. For example, let consider Table 3.3 where the columns represent different types of social data (e.g. *Friend*, *Interest*, *Post*, *Following Post*) and the rows represent different social accounts (e.g. Facebook, Twitter, LinkedIn). The user can freely choose which social accounts along with which types of social data to share with the group. The only rule is that the user has to open at least the *Post* social data of one of his/her social accounts.

TABLE 3.3: Example of a member's sharing settings: the light-gray color means that the element has to be shared by default

Account \ Social Data	Friend	Interest	Post	Following Post
FB	no	no	no	no
TW	no	no	yes	yes
LI	no	no	yes	yes

The user, as indicated in Table 3.3, shares his Twitter and LinkedIn accounts. Consequently, the posts from these two social accounts will by default be selected to match with the group's topics of interest in order to extract the contents of interest. The user moreover decides to also share the following posts with the group. Other social data like friends and interests will not be disclosed to the group.

The third attribute called "review" is optional and complementary to the two first ones. The user has the ability to review every piece of detected relevant content before sharing, thus deleting sensible information. This option can furthermore be used as a collaborative filter to filter out "false positive" information that automated methods missed.

3.3.2   Collective & Personalized Interests

As mentioned above (see *Sharing requirements*), a group's topics of interests should be collectively and dynamically defined by any of its members. In other words, any member is able to propose a new topic and/or suggest additional selectors associated with certain topics whenever he/she finds them relevant to the group. For example, within a group interested in politics, a member can at a given moment suggest a new topic about the upcoming important political event (e.g. a presidential or local election) so that the group can start to capitalize the event-oriented information and news. Such collective principle allows the group to benefit from the expertise of each of its members, thus having appropriate topics and precise selectors.

However, this collective way of defining the group's topics of interest may lead to an important number of topics. It is unlikely that all topics fit the needs of all members. A given member may personally find some topics too broad or too specified or even unsuitable, and may want to ignore them. With this in mind, we have therefore included the features offering each member the ability to personalize his/her interests towards the group's collective interests. More specially, the member does not need to accept all proposed topics, but can follow only a subset of topics that interest him/her the most, also unfollow them later if he/she want to.
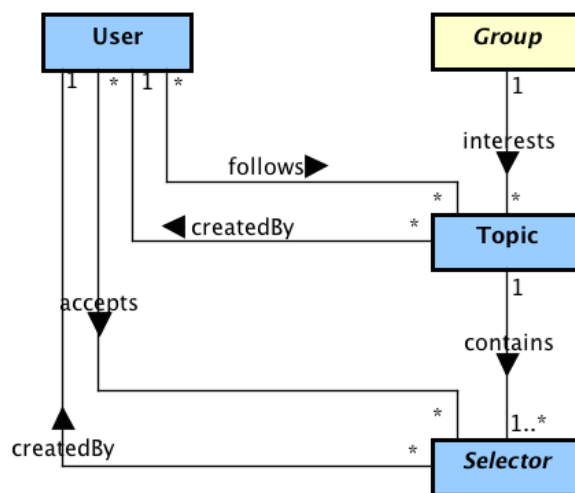
FIGURE 3.7: Collective and Personalized Interests within a group

Following a topic implies default acceptance of all of its current selectors, but the member can later deselect certain selectors if he/she wants to (see Figure 3.7). The member can moreover suggest new selectors to the topic. Every time a new topic or an additional selector is added, it will be spread to other members so that they can decide whether to accept or to ignore it according to their own preferences.

These personalization features have twofold purpose. Firstly, they prevent the members from facing the overload of topics, and subsequently the overload of contents of interest when visiting the group. Secondly, they provide the group with simple means to measure the relevancy of each topic and selector. The more a topic or a selector is followed by the members, the more relevant it is.

Along with a member's sharing settings, his/her personalized topics will be taken into consideration during the filtering process. Only authorized social data will be selected for matching against only the selectors that the user has accepted. Non-matching contents will never be shared with the group.

3.4 Summary

We have presented in this chapter our conceptual answer to the three questions Q_1, Q_2, Q_3, which is essentially composed of two main components *User-centered social data filtering* and *Group-based social data sharing*. With the conceptual foundations detailed above, these two components provide the users with many significant advantages.

The *user-centered social data filtering* component answers the question Q_1 by giving the user a centralized access to interesting contents extracted from the social data aggregated from his/her different social networks. By organizing the contents of interest by topics, it furthermore allows the user to split their large social streams into a number of topic-related sub-streams. This way, the user can easily access to the expected contents by selecting the corresponding topic. Most importantly, the fact that the topics are explicitly provided and are not restricted to a given domain, would allow the user to better exploit the social data which are very various and constantly changing.

The *group-based social data sharing* component answers the questions Q_2 and Q_3. It allows a group of interest to tap into its members' social data without any extra effort, thus increasing the number of information sources (Q_3). Every member can be an active contributor even if not necessarily active in publishing contents in SNSs, since he/she can also share the contents published by his/her social friends. Both the contents published by the members and the contents published by their social friends are "reliable" information sources based on two assumptions: (1) a person who is interested in a topic often tries to share interesting information or useful content in order to influence his/her friends (i.e. *social influence* [100]), (2) people with similar interests are more likely to be connected (i.e. *homopholy* [104]). Taking part in a group of interest, a user can therefore access to more useful contents (Q_2).

On the other hand, the collective principle of defining the topics of interest of a group encourages its members to contribute their expertise. They can suggest a new topic as early as it becomes a trending topic in SNSs, or add more precise, advanced selectors to improve the filtering precision. Moreover, by accepting or rejecting a topic or a selector, the members can promote or demote it. In the second case, the topic or selector in question should gradually disappear from the group. So, if the members are active enough, the group will have enriched, updated topics and precise selectors, and thus end up with more appropriate contents.

It is important noting that the conceptual design described through this chapter remains generic, as it does not impose any technical choice yet. Different technical alternatives can accordingly be proposed. In the next chapter, we will present our technical solution as an instance of this conceptual design.

FIGURE 3.8: Overall Modelling

Chapter 4

Technical Solution

Contents

In the previous chapter, we have seen the conceptual design of our answer to the three questions that we have raised at the beginning of this thesis. It now remains to define the technical means for achieving this design and turning it into reality. As mentioned above, there may be several possible technical implementations. We will present our implementation in this chapter.

More specially, we propose a centralized modular system architecture composed of three major components (see Figure 4.1):

1. Aggregating component,

2. Searching component,
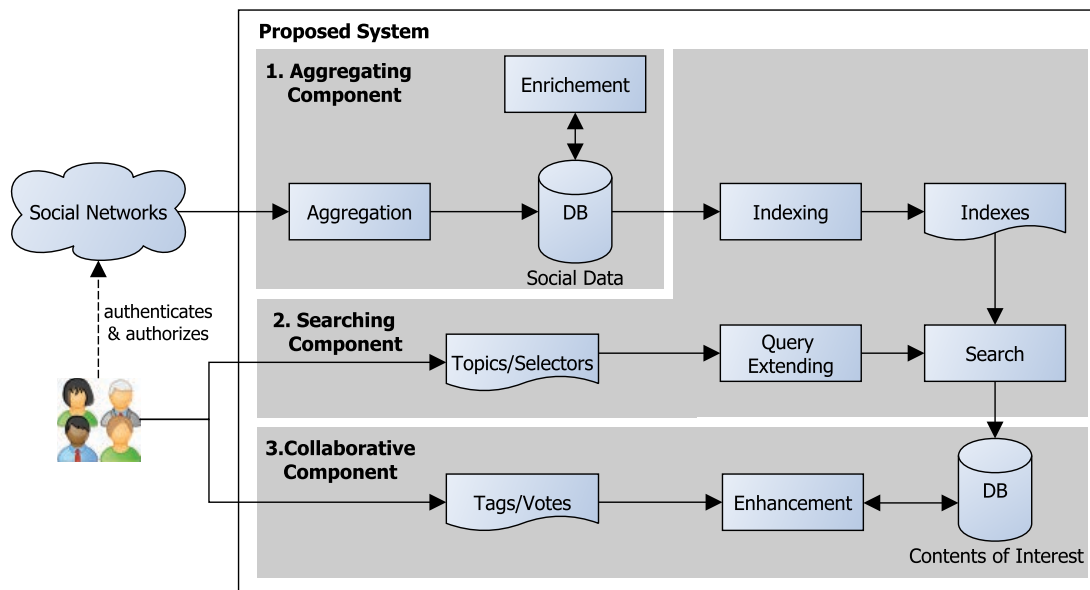
FIGURE 4.1: Proposed system architecture

3. Collaborative component.

These three components have the specified roles and functions within a repetitive process, in which the aggregating component starts to aggregate and store the users' social data, which are therefore enriched. Next, the searching component steps in to index the enriched social data and to launch the personalized searching tasks taking into consideration the users' topics and selectors in order to extract contents of interests. Finally, the collaborative component, based on the group members' collaborative efforts, enhances the quality of the detected contents of interest. This process is repeated at regular intervals to make sure that the users' recent social data are continuously aggregated from different SNSs, processed and filtered.

To make sure that such a process works correctly, each of the three technical components contains a number of specified modules. Below, we will detail the role of each of these modules, its current implementation as well as its eventual issues and possible improvements.

4.1   Aggregating Component

The aggregating component is technically the most straightforward component. It includes three main modules (see Figure 4.2), (1) *Social data aggregation*, (2) *Social data storing*, and (3) *Social data enrichment*, which are responsible for retrieving the users' social data from different SNSs, storing and enriching them, respectively.
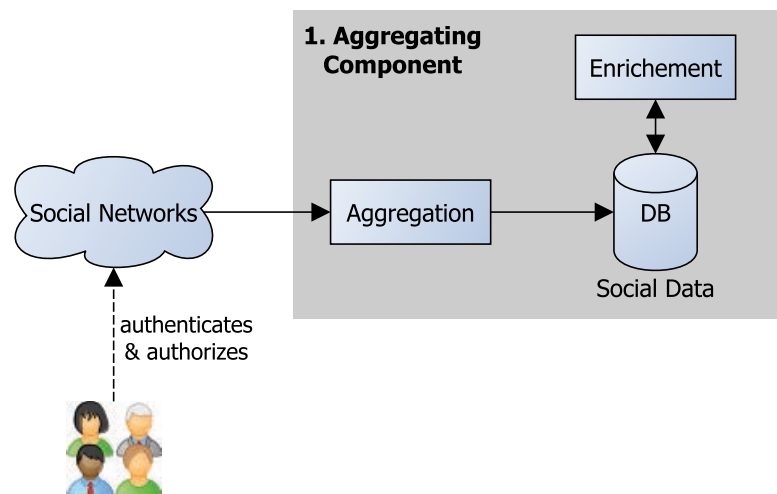
FIGURE 4.2: Aggregating Component

4.1.1 Social Data Aggregation

Among the methods discussed in the two subsections *Unique User Identification* and *Social Data Collection* of Chapter *Literature Review*, relying on the authentication protocols provided by the SNS providers and their proprietary APIs is actually the most suitable solution to our approach for aggregating the users' social data. There are two reasons for it. First, we intended to experiment our proposed approach on some popular SNSs such as Facebook, Twitter or LinkedIn, which unfortunately do not support open standards yet. Second, we would like to comply with the different privacy policies imposed by the SNS providers, and most importantly to have a full access, granted by the users, to their social data.

Thereby, our *Social data aggregation* module is composed of a number of *aggregators*. Each of them is dedicated to a particular SNS for aggregating the user's social data. Obviously, the user first has to authenticate his/her different social accounts on different SNSs, and to grant the aggregators an access to each of these accounts using the dedicated interfaces. With the granted permissions, the aggregators will then be able to request the different APIs (e.g. Facebook Graph API[1], Twitter Rest API[2]) for collecting the users' recent social data at any time.

Actually, we have not created ourselves the different aggregators but adapted the open source library HybridAuth[3], which is delivered with a number of specific classes already including codes for dealing with the different authentication protocols and APIs imposed

---

[1]Facebook API : https://developers.facebook.com/docs/graph-api
[2]Twitter API: https://dev.twitter.com/docs/api/1.1
[3]HybridAuth homepage: http://hybridauth.sourceforge.net/

by the different SNSs providers. Hence, we only needed to extend the existing classes by adding to each of them five methods (i.e. `getUserProfile()`, `getUserFriends()`, `getUserPosts()`, `getUserFollowingPosts()`, `getUserInterests()`). Each method contains a set of *hand-crafted mapping rules* indicating which social data to be requested. Only the social data corresponding to the entities defined by our previous social data model is retrieved by the aggregators.

---

input: The list of users U of size n

```
1  for i ← 1 to n do
2      user ← U[i];
3      socialAccounts ← GetSocialAccounts(user);
4      for j ← 1 to SizeOf(socialAccounts) do
5          account ← socialAccounts[j];
6          token ← GetToken(account);
7          if token ≠ NIL then
8              originId ← GetOriginId(account);
9              socialNetwork ← GetSocialNetwork(account);
10             /* We request only the most recent social data since the
                  last updating time.  If it is the first time, updateTime
                  will be the last two weeks.                            */
11             updateTime ← GetUpdateTime(account);
12             socialData ← Request(socialNetwork,originId,token,updateTime);
13             mappingRules ← GetMappingRules(socialNetwork);
14             if socialData ≠ NIL then
15                 if socialData = ERROR then
16                     token ← NIL;
17                 else
18                     InsertOrUpdate(socialData,mappingRules);
19                     SetUpdateTime(account);
20                 end
21             end
22         end
23     end
24 end
```

Algorithm 1: Aggregation Algorithm

The aggregation process is carried out as described by the Algorithm 1. In brief, it first takes as input the list of all users, and gets the social accounts of each of them. Then, for each account, according to its origin (i.e. the social network), the suitable *aggregator* is launched with a number of parameters including the encrypted permissions (i.e. *token*), the original identifier of the account, and especially the last request time to discard the already requested data. Afterwards, the new social data returned by the corresponding API, if any, are mapped to the underlying model before being stored.

The disadvantage of this technique 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 currently collected at regular time intervals (i.e. *polling*). 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 (i.e. *event-driven*).

4.1.2   Social Data Storing

The *Social data storing* module is responsible for storing social data collected by the *Social data aggregation* module. For that purpose, we have considered two types of databases, namely SQL databases and RDF databases. RDF databases offer a standardized storage solution with a simple, uniform, schema-less data model and a powerful, declarative query language (i.e. SPARQL). RDF databases are often recommended instead of SQL databases when dealing with the data portability and the interoperability among different databases. Nevertheless, SQL databases, especially open source databases, are more popular and well documented than RDF databases, at least for now. Various packages/libraries have furthermore been proposed to greatly facilitate the development over the SQL databases.

For practical reasons, we have at the moment implemented the social data storing module with a SQL database, namely MySQL (see the physical schema in Appendix MySQL Physical Schema). This option allowed us not only to quickly set up a reliable database, but also to reduce the development time of our first Web-based prototype that we will see in the next chapter. On the other hand, as our underlying data models (i.e. social data integration model and group-based content organization model) contain a small number of entities and relationships, the corresponding relational database schema remains simple, thus being in principle efficient and quick in terms of data insertion and request.

The *scalability* and *flexibility* criteria do not at this stage play an important role, but will become critical when the number of supported social networks, the number of users and subsequently the amount of social data dramatically increase. Likewise, the *data portability* and the *interoperability* factors may be also required within some of the future advanced use scenarios of our proposed approach. To meet these criteria, a RDF database should be obviously considered to substitute the current relational database.

4.1.3   Social Data Enrichment

The *Social Data Enrichment* module attempts to enrich the textual content of the aggregated social data. This step is necessary to improve the effectiveness of the subsequent

filtering step, as social data often contain very little text and are ungrammatically written. We have applied one technique that mostly concerns the social data containing external links. It consists of expanding the textual content of the social data with the additional content extracted from the referred web pages (e.g. the titles and the descriptions). Despite its simplicity, this technique is very helpful, since lots of social data contain links to external Web resources.

In addition to this enrichment technique, we could apply other advanced methods of natural language processing (NLP) [159] to enrich the social data. For example, the *language detection* step could be added to determine the language, in which the textual content of the social data is written. Given the language, the subsequent modules of the system would be able to process the social data in a more in-depth manner. The *entity extraction* techniques could also be applied to extend the social data with the descriptions and/or the categories of its containing entities. However, such advanced helpful techniques would require a considerable execution time taking into consideration the big number of social data to process. A efficiency test on a given sample of social data would therefore be necessary before deciding to apply one of these advanced enrichment techniques.

4.2   Searching Component

As we have seen in the section *Information Filtering* (IF) of the chapter *Literature Review*, there is a variety of model-driven methods proposed for filtering social data. These methods are dedicated to a specific SNS, or a specific domain, or even a given language. Although, they are efficient within their application scope, they are hardly able to fit the social data about other domains or from other SNSs. Some of them moreover require regular training, unless they will become obsolete towards the social data, which constantly evolve and appear in a huge number.

Taking it into consideration, we have applied other generic techniques originated from the *Information Retrieval* (IR) area which is closely linked to the IF area [18]. IR is aimed at finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [99]. Technically, the documents are retrieved, upon search queries specified by users, based on meta-data or on full-text indexing of the documents (not the documents themselves). For that matter, we preferred the term "searching" rather than "filtering" to name the component responsible for extracting the contents of interest.
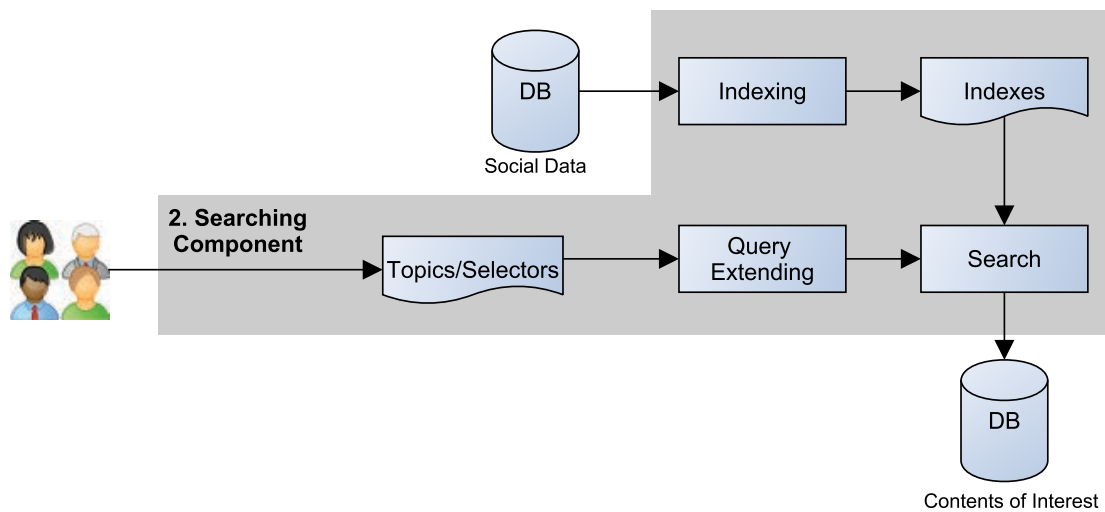
FIGURE 4.3: Searching Component

Moreover, we have chosen to implement this searching component using the open source Lucene platform[4]. This choice, inspired by a number of successful Lucene-powered social data analysis works [66, 116, 117, 125], offers numerous advantages. We do not need to develop our own search engine, and most importantly Lucene is widely approved for providing a robust and scalable indexing and retrieval platform that is designed to cope with Web-scale data and usage [103].

Since Lucene is a text-based search engine, its basic units of information are documents, which are indexed and stored for retrieval. We therefore treat the users' enriched social data as documents and their topics of interest as information need. Especially, in our case, the searches are automated in accordance with the filtering requirement. The users explicitly express their topics of interest in terms of special queries. The system regularly searches for social data matching these queries as long as the new social data are indexed.

We detail below how the social data are indexed, how the users can specify their topics of interest, and how the contents of interest are retrieved, respectively.

4.2.1 Social Data Indexing

The *Social data indexing* module is necessary to generate for each new piece of aggregated and enriched social data its corresponding indexes, which will be later used for the retrieval task. The indexing process as illustrated in Figure 4.4 consists of first listing the different fields to analyse and/or to index, then transforming the field values into index terms which are then written in an inverted index table.

---

[4]Lucene: http://lucene.apache.org/java/docs/
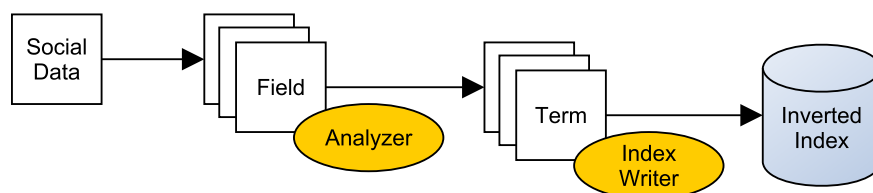
FIGURE 4.4: Indexing Process

Field	Description	Analysed	Indexed	Stored
id	the ID generated by the system when inserting the social data	No	Yes	Yes
timestamp	the created time of the social data	No	Yes	No
type	the type of the social data (e.g. "friend", "interest", "post", "following post")	No	Yes	No
social network	the source of the social data	No	Yes	No
owner	the user ID who own the social data	No	Yes	No
text	the text content of the social data plus the enriched part	Yes	Yes	No

TABLE 4.1: The indexed fields of the social data

With respect to the previously defined social data model (see Figure 3.8), we have declared a number of fields to be indexed including *id, timestamp, type, social network, owner* and *text*, each of which plays a specified role that we will explain further later in this section. Moreover, as shown in Table 4.1, they are treated differently. Only the *id* field is stored with the original value in order to retrieve the original social data. Also, only the *text* field needs to be analysed before being indexed.

To analyse the *text* field, we have chosen the standard analyzer (i.e. `StandardAnalyzer`) of Lucene. It is a general-purpose but quite sophisticated analyzer which is able to:

1. Tokenize the text content of the social data, which means that it breaks down the initial text into words using the whitespaces and the common delimiters;

2. Lowercase each word to make it non-case-sensitive;

3. Remove stop words that are high frequency words like "the", "a".

This standard analyzer does the job fairly well for various western languages like English, French, or Spanish. Nevertheless, to support more languages and to improve the analysis
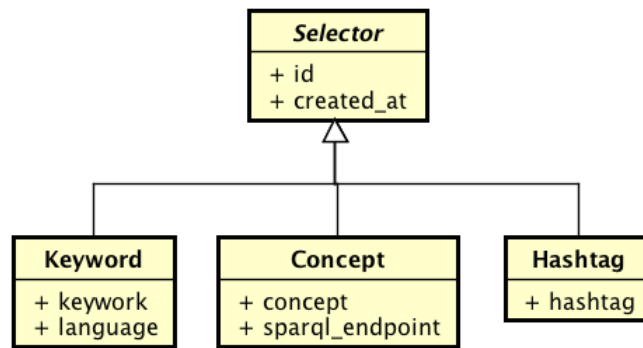
FIGURE 4.5: Selector types

quality, we could add an extra step that detects the language of the text, and then use a language-specific analyzer (e.g. `GermanAnalyzer`, `FrenchAnalyzer`, `SpanishAnalyzer`).

4.2.2 Developed Selectors

As mentioned in the last chapter, we would like to set up a topic-selector structure for organizing and extracting contents of interest. The *topic* level corresponds to the topics of interest, whereas the *selector* level represents the technical specifications of *topics*, which give guidance on how to do the retrieval process. We have, at this stage, developed three different types of selectors, namely *hashtag-based* selector, *keyword-based* selector, and *concept-based* selector (see Figure 4.5) that the users can freely choose to specify their respective topics. Each type of selector has its proper characteristics as follows:

- *Hashtag-based* selectors expect a valid *hashtag*, which is a word or a phrase prefixed with the symbol "#", as value. Hashtags are widely adopted by social network users to collectively group and efficiently retrieve their messages [109]. Likewise, our hashtag-based selection is also an "upstream" effort. When posting some content on a given SNS, the users can include a previously chosen hashtag to explicitly indicate its relevancy. So, contents containing such a hashtag will be directly selected as contents of interest.

- *Keyword-based* selectors follow the same principle of web search query, thus accepting either a single word or several words combined by boolean operators such as "OR", "AND", "NOT" [99] as value. Furthermore, the language of the keyword, if provided, will allow to expand the initial keyword with its derived forms and/or its synonyms using dedicated dictionaries (e.g. WordNet).

- *Concept-based* selectors require a referenceable concept belonging to an ontology publicly accessible via a SPARQL[5] query endpoint[6]. The users, in particular when belonging to a group of interest, are encouraged to use their domain-specific ontology. Otherwise, they may already use other open and multi-domain ontologies like DBpedia[7], which is a generic knowledge base containing millions of multi-language and multi-domain entities [22]. Compared to keyword-based selectors, *concept-based* selectors are more powerful. Firstly, they provide multi-languages labels for a single concept. Secondly, they allow to expand the given concept by its related concepts, thus disambiguating the concept and improving the matching precision as well.

With the three types of selector, we expect to give the users various useful choices ranging from collaborative selection to domain-specific selection. The two user-friendly methods, hashtags and keywords, do not require users to have specific knowledge. The concept-based selectors enable an advanced selection of contents of interest. The user can set and assign as many instances of one of the three types of selector as desired to a given topic.

It is important to note that this list of selectors is not definitive and may be supplemented by other types of selector. For example, some heuristic filters can be included and applied to the output of the aforementioned selectors to remove some too short or too personal contents (e.g. contents about *me now, presence maintenance, anecdote* [110]).

4.2.3   Query Expansion

The *Query expansion* module is responsible for translating a selector entered by a user into a internal query. It furthermore expands the given value of the selector with additional values according to its type.

For hashtag-based selectors, there is no need for expansion. The only thing to do is to preserve its "#" symbol from being tokenized by the internal text analyzer. This is done by temporarily replacing it by a text-based value, for example "HT".

In the case of a keyword-based selector, the *Query expansion* module, whenever it is provided with a single word, will expand the given word with its derived forms and possibly its synonyms according to the language of the word. For that purpose, it is based on a number of dictionaries from the Python module called Pattern[8]. This module now supports six different languages (i.e. English, Spanish, German, French, Italian, and Dutch), and provides tools for verb conjugation and noun singularization and pluralization. Especially,

---

[5]SPARQL: SPARQL Protocol and RDF Query Language is an RDF query language
[6]The system does not have importing, reading, parsing and version managing features for ontologies yet
[7]DBpedia: http://dbpedia.org/About
[8]Pattern homepage: http://www.clips.ua.ac.be/pattern

it contains a WordNet[9] interface for looking up the synonyms of an English word. After gathering the variants of the given word, the module will build the final query by concatenating these variants and the given word using the "OR" operator, for example "*automobile OR automobiles OR car OR auto*".

If the keyword-based selector is set with several words linked by boolean operators, the expansion is a bit more complicated. The module first needs to locate the containing words, then replace them by their corresponding *OR-concatenations*. Here below are three examples of combined keyword-based selectors (q_{or}, q_{and}, $q_{and \oplus or}$) and their expanded forms (q'_{or}, q'_{and}, $q'_{and \oplus or}$):

$$q_{or} = k_1 \vee k_2 \rightarrow q'_{or} = k_1 \vee k'_1 \vee k''_1 \vee k_2 \vee k'_2 \tag{4.1}$$

$$q_{and} = k_1 \wedge k_2 \rightarrow q'_{and} = (k_1 \vee k'_1 \vee k''_1) \wedge (k_2 \vee k'_2) \tag{4.2}$$

$$q_{and \oplus or} = (k_1 \wedge k_2) \vee k_3 \rightarrow q'_{and \oplus or} = ((k_1 \vee k'_1 \vee k''_1) \wedge (k_2 \vee k'_2)) \vee k_3 \tag{4.3}$$

Where k_1, k_2, k_3 are three keywords, and k'_1, k''_1 are the variants of k_1, and k'_2 is the variant of k_2.

For concept-based selectors, the Query expansion module currently requests the SPARQL endpoint for the (multi-language) labels of the given concept. For example, the concept *dbpedia:automobile* provides "automobile", "automobil", "automóvil" as labels, and subsequently leads to the query "*automobile OR automobil OR automóvil*". Concept-based selectors could be further semantically expanded if we would associate the given concept with its closely related concepts in order to disambiguate it. For example, the category "city" may be appended to "Paris" (i.e. *"Paris" AND "city"*) to make sure that the results should be related to the capital of France.

As mentioned above, a user can assign as many selectors as desired to a topic in order to increase the probability of retrieving contents of interest. Let's take the example of the topic "Automobile" illustrated in Figure 4.6 where it is associated with three selectors: a hashtag-based selector (i.e. "#automobile"), a keyword-based selector (i.e. "automobile"), and a concept-based selector (i.e. "dbpedia:automobile"). For each of the three selectors, the Query expansion module will generate the corresponding final queries.

It is worth noting that the queries correspond to a technical (low) level and are created at runtime. For this reason, we have not include it in our aforementioned three-level organization structure (i.e. Group - Topic - Selector).

---

[9]Wordnet is a lexical database that groups related words into Synset objects (= sets of synonyms) http://wordnet.princeton.edu/
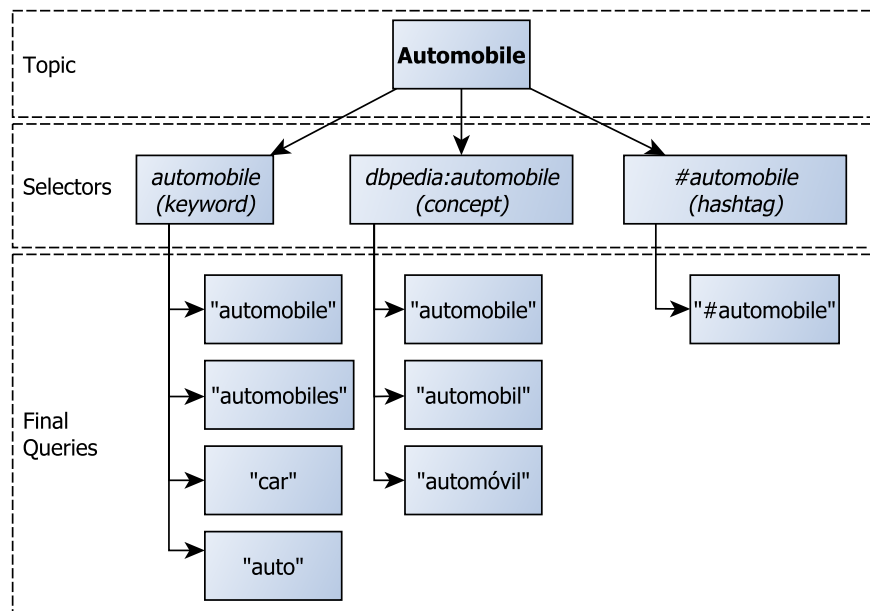
FIGURE 4.6: Example of query expansion

4.2.4   Content Searching

The *Content searching* module is responsible for retrieving contents of interest, and is executed per group. The whole process is described by Algorithm 2. In short, the module first gets the list of all the groups with their respective member list. For each member of a given group, the module creates as many queries as selectors that the member has chosen taking into consideration his/her sharing settings and personalized topics of interest. The final queries are then searched against the index of all aggregated social data. The top retrieved results will be saved as contents of interest and associated to the corresponding topics and groups.

Let's take a deeper look into the algorithm. Line 13 is added to get the expanded form of the selector following the principles provided in the last subsection. The expanded query is at this stage not complete. The `BuildFinalQuery()` function (Line 14) transforms it into a final query taking into consideration the user's sharing settings, by including the additional fields (see Table 4.1) in the following order:

1. Add the authorized social data types, for example, `type:"Post OR FollowingPost"`,

2. Add the authorized social accounts, for example, `social network:"Facebook OR Twitter"`,

3. Add the owner, for example, `owner:"User1"`,

```
input: The list of groups G of size n
input: The most recent index I of social data
input: The last searching time T
input: The maximal number of retrieved contents k
1  for i ← 1 to n do
2  │  group ← G[i];
3  │  members ← GetMembers(group);
4  │  for j ← 1 to SizeOf(members) do
5  │  │  user ← members[j];
6  │  │  sharingSettings ← GetSharingSettings(user,group);
7  │  │  authorizedTypes ← GetAuthorizedTypes(sharingSettings);
8  │  │  authorizedAccounts ← GetAuthorizedAccounts(sharingSettings);
9  │  │  review ← GetReview(sharingSettings);
10 │  │  selectors ← GetSelectors(user,group);
11 │  │  for k ← 1 to SizeOf(selectors) do
12 │  │  │  selector ← selectors[k];
13 │  │  │  expandedQuery ← GetExpandedQuery(selector);
14 │  │  │  finalQuery ← BuildFinalQuery(expandedQuery, authorizedTypes,
   │  │  │    authorizedAccounts, user, T);
15 │  │  │  contents ← Search(finalQuery, I, k);
16 │  │  │  if contents ≠ NIL then
17 │  │  │  │  topic ← GetTopic(selector);
18 │  │  │  │  for l ← 1 to SizeOf(contents) do
19 │  │  │  │  │  content ← contents[l];
20 │  │  │  │  │  Save(content, group, selector, topic, review);
21 │  │  │  │  end
22 │  │  │  end
23 │  │  end
24 │  end
25 end
26 T ← Now();
```

Algorithm 2: Searching Algorithm

4. Add the time constraint, for example, `timestamp:[Last 24 hours,Now]` (from the last 24 hours).

And it produces a query like:

```
owner:"User1" type:"Post OR FollowingPost" timestamp:[Last 24 hours,Now]
social network:"Facebook OR Twitter" text:"TheExpandedQuery"
```

This final query is searched against the most recent index of all aggregated social data in order to retrieve the *top-k* most relevant contents (using the function `Search()` in Line 15). Actually, the contents are selected through two steps: first, a subset of social data which meet the added conditions (i.e. the type, social network, owner, and timestamp fields), is

extracted, then, these social data are matched against the expanded query (i.e. the text field).

For the second step, we have used the *Extended Boolean Model* [127] natively integrated in Lucene. This model combines the characteristics of the Vector Space Model (VSM) [128] with the properties of the Boolean Model (BM) [91]. It first uses the BM to narrow down the list of documents (i.e. hits) that need to be scored based on the use of Boolean logic in the query specification, then uses VSM to determine how each of them is relevant to the query (i.e. scoring). This way a document may be somewhat relevant if it matches some (not all) of the queried terms and will be returned as a result.

In VSM, documents and queries are represented as weighted vectors in a multi-dimensional space, where each distinct index term is a dimension. The VSM score of a document d for a query q is the *Cosine Similarity* of their weighted vectors $V(q)$ and $V(d)$:

$$cosine\_similarity(q, d) = \frac{V(q) \cdot V(d)}{|V(q)||V(d)|} \qquad (4.4)$$

where $V(q) \cdot V(d)$ is the scalar product of the two weighted vectors, and $|V(q)|$ and $|V(d)|$ are their *Euclidean norms*. For search quality and usability, Lucene refines this VSM score and derives a *practical scoring function*[10] using TF-IDF weighting [128]:

$$score(q, d) = coord(q, d) \cdot queryNorm(q) \cdot \sum_{t \in q}(Tf(t, d) \cdot Idf(t)^2 \cdot t.getBoost() \cdot norm(t, d))$$

$$(4.5)$$

where

- $Tf(t, d)$ correlates to the term's frequency, defined as the number of times that the term t appears in the currently scored document d. Note that $Tf(t, q)$ is assumed to be 1 and therefore does not appear in this equation. However, if a query contains twice the same term, there will be two term-queries with that same term. $Tf(t, d)$ is computed as follows:

$$Tf(t, d) = frequency^{1/2} \qquad (4.6)$$

---

[10]Lucene TFIDF Similarity Formula: https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

- $Idf(t)$ stands for *Inverse Document Frequency*. Its value correlates to the inverse of the number of documents in which the term t appears. This means that rarer terms give higher contribution to the total score. $Idf(t)$ appears for t in both the query and the document, hence it is squared in the equation 4.5. $Idf(t)$ is computed as follows:

$$Idf(t) = 1 + ln(\frac{|D|}{|\{d \in D : t \in d\}| + 1})$$ (4.7)

- $coord(q, d)$ is a score factor based on how many of the query terms are found in the specified document. Typically, a document that contains more of the query's terms will receive a higher score than another document with fewer query terms.

- $queryNorm(q)$ is a normalizing factor used to make scores between queries, and does not affect the document ranking.

- $t.getBoost()$ is the boost value of the term specified in the query. In our case, all terms are equal, no term is boosted.

- $norm(t, d)$ is a normalization factor for the document length, more precisely the field length. It is in accordance with the number of tokens of the field which contains the term t. In principle, shorter fields contribute more to the score.

We have not planned at this time to override this default scoring formula of Lucene. Nevertheless, it remains to define the maximal number of retrieved contents k for a given query q ($k \geqq$ the number of contents to save). This is not a trivial task given that k depends on several factors such as the complexity of the expanded query (C_q), the user's sharing settings (S_u), the total number of hits ($H_{u,q}$), and so forth. Thereby, we have identified four different ways to define the value of k:

1. Get all of retrieved contents, so there is no need for specifying k, and $k \geqq 0$;

2. Fix the value of k, often as a small value for example $k \in \{5, 10, 20\}$;

3. Compute k at search time taking into consideration the aforementioned factors, $k = f(C_q, S_u, H_{u,q}, ...)$;

4. Combine these methods taking into consideration the number of members in the group, the value of $\sum_{u \in U'} H_{u,q}$ where U' is the set of members who have chosen the same selector. For example, if the group is private or small or there are too few retrieved documents, then get all of them. If there are too many candidate contents, then either limit k at a small number (e.g. 10) or dynamically compute k for each user u.
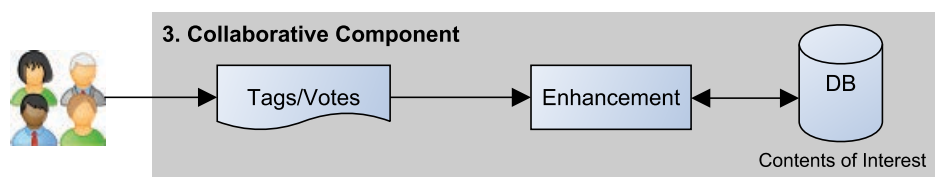
FIGURE 4.7: Collaborative component

At the moment, we have applied the second strategy, and fixed k at 10. We have furthermore set up 4 searches at various times of the day which means for a query, there may be at maximum 40 contents per day. This static way for valuing k is quite good to the current scope of the system where there are mostly small groups. Later, when a group grows in size and contains some very generic and/or popular queries, the number of its contents of interest may increase considerably. Therefore, it would be interesting to explore further other methods, in particular 3 and 4, and apply one of them to make the retrieval module more selective.

4.3   Collaborative Component

The *Collaborative* component covers advanced features mostly concerning the open groups where we can encourage and benefit from the collaborative contributions of all members (see Figure 4.7). At this stage of the project, there is only one integrated module called *Enhancement* that we will detail below. Other interesting modules to study will be discussed in the subsection *Group-specific Knowledge Discovery* of the chapter *Perspectives and Future Work*.

4.3.1   Enhancement

The Searching component, in particular its content searching module, has currently been implemented with the basic configuration of Lucene. The *top-k* of retrieved contents is furthermore fixed at a relatively high value (i.e. 40 contents per day per group member). All of this may cause a gap of the searching performance (e.g. *false positives*). The *Enhancement* module is therefore very useful, as it makes it possible to improve the quality of a group's contents of interest using its members' collaborative efforts. More specially, the members are given several practical means of contribution as follows:

1. A member can enable the *review* option (see the subsection *Group Settings* of the chapter *Conceptual Design*). So, each time the member visits the group, he/she is

notified about the newly detected contents, if any, and should take appropriate actions (i.e. validate or ignore each of these contents). This way, the member can remove sensible information, and also spot the false-positive contents which were retrieved by the retrieval module but are not really interesting.

2. Even if the member does not enable the *review* option, which means that all detected contents are immediately shared with the group, he/she still can delete the contents belonging to him/her whenever he/she wants to.

3. For each content of interest of the group, the member can vote it as "relevant" or "irrelevant" to respectively promote or demote it. A content with more relevant votes can be highlighted to draw more attention of the other members. In contrast, when it receives a certain number of irrelevant votes (V), it will be definitively removed from the group.

4. In addition to voting, the member can tag a content with an additional topic if he/she thinks it appropriate. If a content is associated with the same topic by a certain number of members (T), it will be officially associated with the topic.

In the points 3 and 4, we have seen two undefined thresholds V and T. Both are not easy to efficiently determine. They depend on many factors such as the size of the group, the expertise of the member who votes or tags a content. We have for now set V and T to some static values, practically at 2. These minimized thresholds are acceptable for small groups with less than 10 members (2/10). Like k (see the subsection *Content Searching*), there is obviously room for improvement.

4.4 Summary

We have shown in this chapter a technical solution in accordance with the conceptual design presented in the previous chapter. This solution has a centralized modular architecture including three main components. They are (i) *aggregating*, (2) *searching* and (3) *collaborative* components. Each of them contains several specified modules implemented with different techniques as summarized in Table 4.2, and has various functions.

The aggregating component is quite straightforward. It constantly aggregates and stores the users' new social data from different SNSs using the provided APIs. Moreover, the social data are enriched by the external resources to improve the performance of the subsequent modules.

The searching component is the most important component of the system, since it is responsible for extracting contents of interest. Taking into consideration the numerous,

multi-language and mostly text-valued natures of the social data, we have implemented this component using the open source Lucene platform, which is known to provide a robust and scalable indexing and retrieval platform. Every piece of social data is indexed so that it is searchable by specific queries which can be keywords or hashtags or concepts. Retrieved social data for a given query are scored and ranked by the default practical scoring formula of Lucene so that the top scored social data will be considered as contents of interest to be shared within interested groups.

The third component is new to the conceptual design, but is in line with the group-based sharing principles. It allows the users to perform and benefit from other more explicit collaborative efforts beyond the abilities of automatically powering a group's collective social data sources and of collectively defining its topics of interest. Furthermore, certain collaborative efforts such as those used by the enhancement module allow to fill the potential gaps of the current searching performance.

It is worth stressing that our modular system architecture with the currently implemented modules defines a baseline technical solution, thus being totally improvable and extensive. As shown in Table 4.2, there is room for improvement in each module. It is also possible to replace a module, or even an entire component, by another more efficient one or to integrate useful additional modules without redesigning the whole architecture.

TABLE 4.2: System module summary

Module	Current Implementation	Possible Improvement
Social Data Aggregation	Based on the APIs provided by the SNSs providers, Use hand-craft mapping rules, Auto-run periodically.	Integrate realtime update features.
Social Data Storing	Use a relational database (i.e. MySQL)	Migrate or map to a RDF database
Social Data Enrichment	Expand the text content with the additional contents from the referred webpages	Apply more sophisticated NLP techniques.
Indexing	Index social data as documents with multiple fields using Lucene	Replace the standard analyzer by language-specific analyzers.
Selectors	Contain three different types (i.e. keywords, hashtags, concepts)	Improve the concept-based selectors by taking into consideration the concept hierarchy as well.
Query Expansion	Expand the query with its derived forms (i.e. plural or singular) and/or its synonyms using the dictionaries of Pattern	Increase the number of supported languages.
Content Searching	Apply the Extended Boolean Model, Use the native scoring function based of TFIDF of Lucene, Fix the value of k (top-k retrieved documents).	Compute k dynamically
Enhancement	Based on collaborative efforts (e.g. manual removal, vote, tag), Fix the value of the thresholds V and T	Compute V and T dynamically

Chapter 5

Web-based Prototype

Contents

We have seen in the previous chapter a baseline centralized system architecture with a number of specified improvable modules allowing to technically accomplish the conceptual requirements of our proposed solution to the three addressed questions. Nevertheless, we have not shown how it can actually be deployed. To this end, there are practically two major approaches:

1. To develop it as an extension of an existing collaborative system which may belong to an organization or an enterprise;

2. To develop it as an independent system.

In the first case, the development should be tailored to the specific needs of the organization (e.g. specified social networks, additional internal policies, etc.). Also, it should be in accordance with the already provided features of the original system. For example, if the existing system does not support group-oriented features yet, we can develop the whole

proposed solution as an extra social layer and make it interoperable with the remainder of the system. Otherwise, we can reuse and extend certain existing components.

In the framework of our research work, we did not seek to deliver a final solution, but to provide a prototype (i.e. proof of concept) which would be operational and accessible for as many people as possible in order to test and evaluate and improve it incrementally. We have thus opted for the second approach and especially developed a Web-based application. This way, the end users are able to access the system from anywhere without restriction.

Our Web-based prototype accessible at[1], is named *SoCoSys* standing for *Social Collective System*, as it allows to aggregate and filter social data and supports collaboration as well. In this chapter, we will first take an overall look at the layered architecture of SoCoSys, then describe its required use cases and its dedicated user interfaces.

5.1   Application Architecture

We have designed SoCoSys using a layered architecture as illustrated in Figure 5.1, which is basically composed of three layers: (1) *User Interface*, (2) *Restful Web Service*, and (3)*Backend Subsystem.*

The User Interface (UI) provides the users with the suitable accesses to the offered features. It is built based on the *responsive web design approach* [49], which is aimed at adapting the presentation of the webpage with respect to the characteristics of the visiting device. The webpage is therefore easy to read and to navigate across a wide range of devices including computers, smartphones, and tablets.

The layer of Restful Web Service acts as an intermediary between the User Interface and the different databases as well as the Backend Subsystem. It receives from User Interface different HTTP-based queries (i.e. GET, PUT, POST, and DELETE), translates them into internal requests, upon which the output is returned to the User Interface using the JSON[2] format. This medium layer is useful for various reasons such as security, performance, modifiability, and reliability.

The Backend Subsystem is exactly the modular system proposed in Chapter *A Technical Solution*, thus containing the same modules. Certain modules of the Backend Subsystem can be activated upon a user request, for example, the Aggregation module is executed when the user connects one of his/her social accounts for the first time, while the other modules are configured with an auto-run feature.

---

[1]SoCoSys: http://212.129.40.98/scs/#/
[2]JSON: JavaScript Object Notation
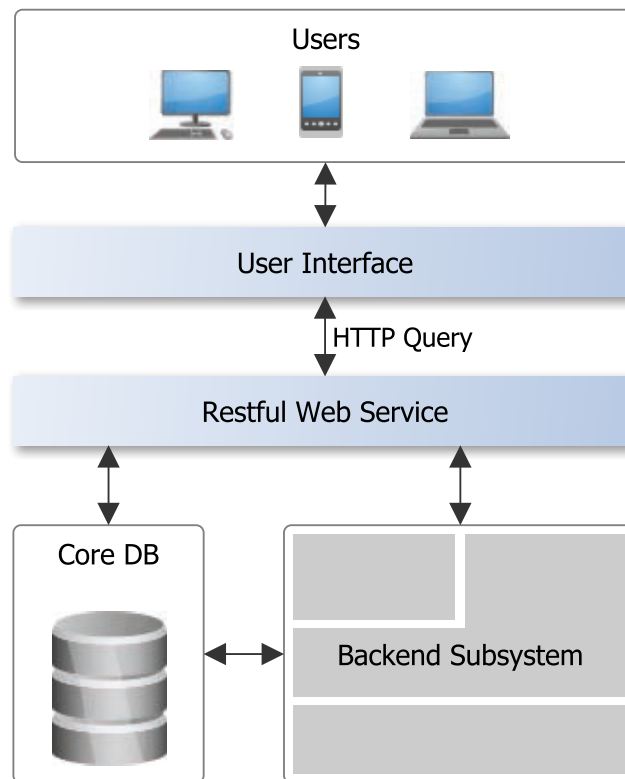
FIGURE 5.1: SoCoSys Architecture

5.2   Use Cases

For building the target prototype, it is very helpful to begin by identifying and describing, at a high level, the required interactions with the system that a user will have to perform using the corresponding functionalities offered by the system to archive his/her goal/objective.

In the case of SoCoSys, the user's main goal is to aggregate his/her social data from different SNSs, to extract from these data the contents of interest, which can then be shared with his/her respective groups. This global objective should be decomposed into a set of specified functions easier for the user to understand and to perform. These functions furthermore drive our incremental development of the target prototype, especially its user web interface.

To define such functions, we have relied on the use case diagrams, which were first introduced by Jacobson and al. in [75]. For ease of reading the diagrams below, we recall the three basic components of a use case diagram, namely *actor*, *use case*, and *relationship* (see Figure 5.2). A use case represents a high level individual functionality of the system. An actor is an external system that interacts with the system for which use case are being created. An actor could be a human being, or any other interfacing system. There are four
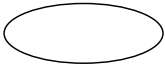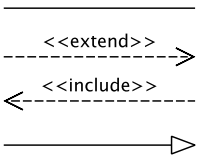
Symbol	Reference Name
	Actor
	Use case
	Relationship

FIGURE 5.2: Symbols of use case diagrams

different types of relationships (see Figure 5.2): the solid line indicates the interaction be-
tween an actor and a use case; the "*extend*" dashed line is between two use cases when one
is an extension of the other under certain conditions; the "*include*" dashed line is between
two use cases when one use case involves and its outcome depends on the resolution of the
other; the solid line terminated by an arrow triangle indicates the relation of generation (or
specification) between two use cases, in which one use case is a particular case of the other.

As illustrated in Figure 5.3, our target system, SoCoSys, has, at the top level, five use cases,
which involve two major actors: *users* and *social network APIs*. A *User* can be either a
New User or a *Registered User*. A *New User* must register with a unique email before
being able to utilize the system. A *Registered User* can perform the general functionalities
like *Manage Social Accounts*, *View Aggregated Social Data*, *Manage Groups*, and *Visit A
Group*.

The *Manage Social Accounts* use case is used by the user to manage his/her different
social accounts. For the moment, SoCoSys supports three SNSs, namely Facebook, Twit-
ter, and LinkedIn. They are the undisputed leaders in their respective domains, which
are general-purpose social networking services, social microblogging services and business-
oriented social networking services. The *Manage Social Accounts* use case is, as shown
in Figure 5.4, extended by three optional use cases *Connect Social Accounts*, *Disconnect
Social Accounts*, and *Reconnect Social Accounts*. They allow the user to connect/discon-
nect/reconnect respectively one or several of his/her social accounts. To these ends, the
three use cases furthermore have to interact with the APIs provided by the SNS providers.
Additionally, like other top-level use cases, the *Manage Social Accounts* use case includes
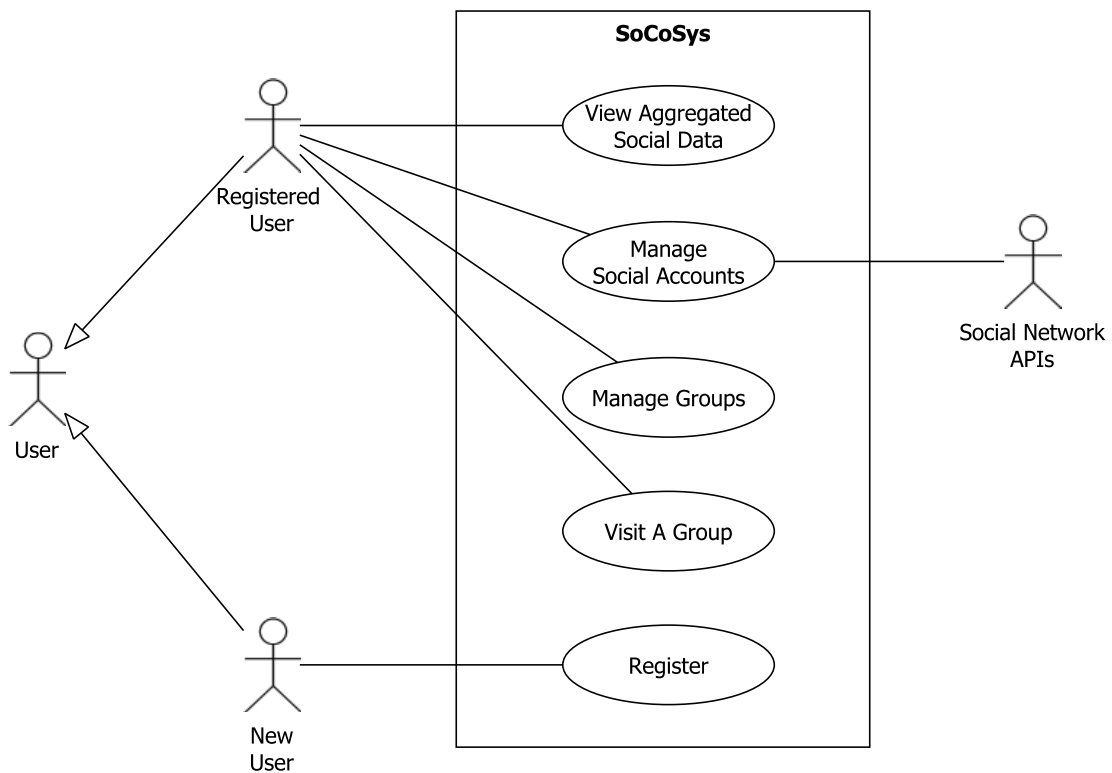
FIGURE 5.3: Top level use cases

a required step, *User Authentication*, which makes sure that all actions are performed by the right person.

The *View Aggregated Social Data* use case allows the user to view the social data aggregated from his/her connected social account(s). The social data can moreover be arranged into different views to ease the browsing task.

The *Manage Groups* use case is aimed at managing the private and open groups to which the user belongs. This use case is extended by three additional use cases *Create A Group*, *Search Groups*, and *Suggested Groups* (see Figure 5.5). The *Create A Group* use case, throughout its two specialized use cases *Create A Private Group* and *Create An Open Group*, allows to create a new private or open group of interest. The *Search Groups* use case allows the user to search for open groups using keywords, while the *Suggested Groups* use case suggests the user new open groups. Especially, during the execution of the *Search Groups* and *Suggested Groups* use cases, the user can decide to join a particular group using the *Join A Group* use case.

The *Visit A Group* use case is the most important one, and should be the most frequently used by the users. It is at least extended by three use cases which are *Edit Sharing Settings*, *Edit Topics/Selectors*, and *View Contents of Interest* (see Figure 5.6). The *Edit Sharing*
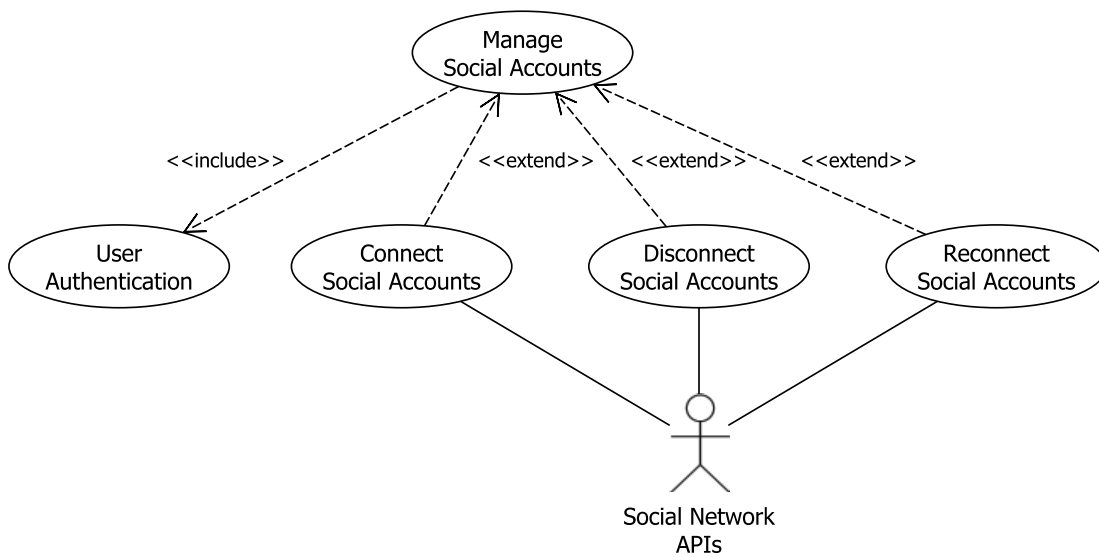
FIGURE 5.4: Manage social accounts use cases
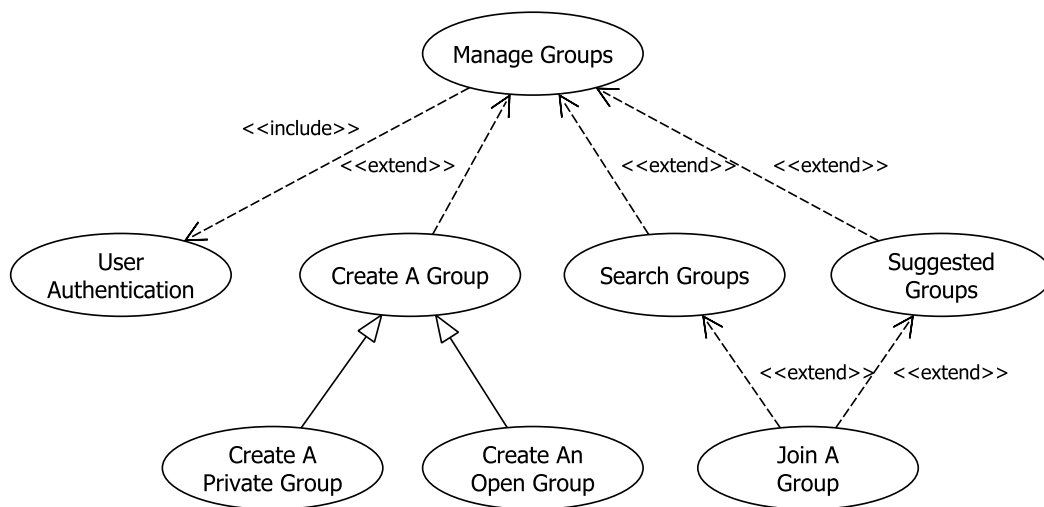
FIGURE 5.5: Manage groups use cases

Settings use case allows the user to adapt his/her sharing settings with respect to the group in question. The *Edit Topics/Selectors* use case, extended by the specific functionalities such as *Add A Topic*, *Follow/Unfollow A Topic*, *Add A Selector*, *Follow/Unfollow A Selector*, allows the user to add a topic, to follow or unfollow a topic, to add a selector, and to follow or unfollow a selector respectively.

The *View Contents of Interest* use case makes it possible for the user to access to the group's contents of interest originated from its collective social data sources. All contents of interest are displayed together within a chronologically ordered stream. Nevertheless, the user can filter this stream by topics using the *Filter By Topic* use case. The user can
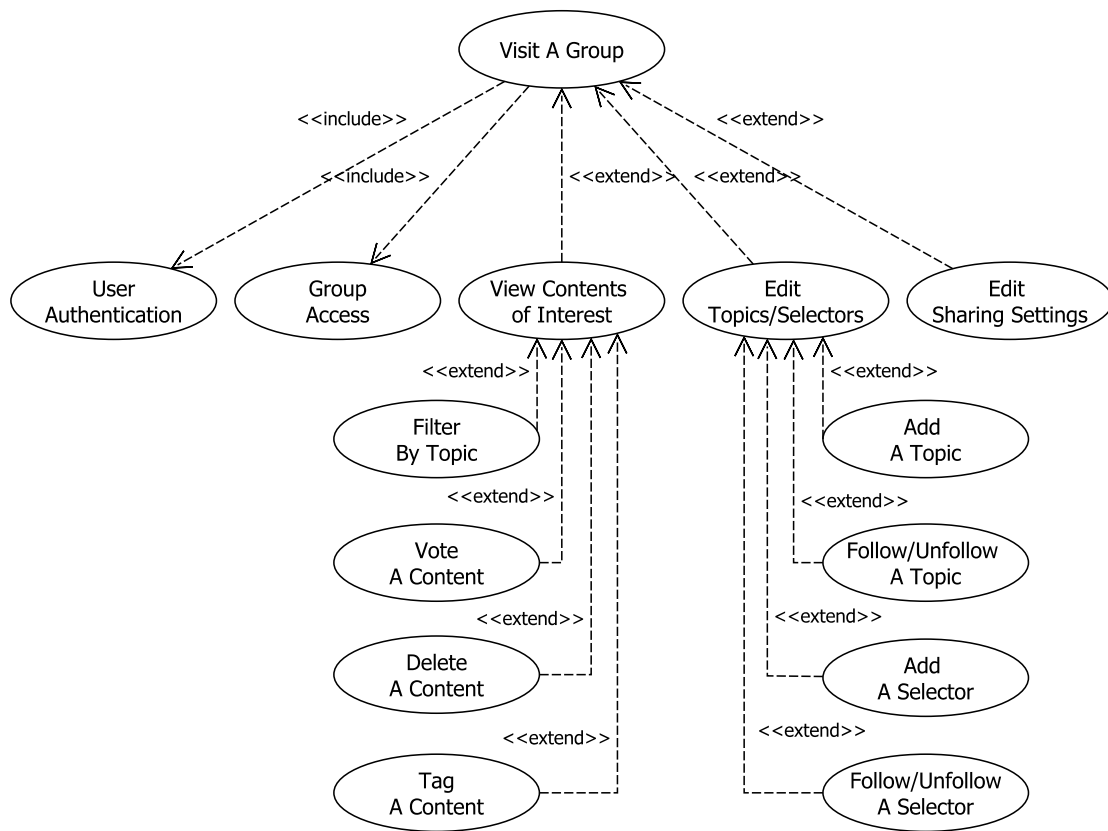
FIGURE 5.6: Visit a group use cases

also vote a content as irrelevant or relevant, delete an unsuitable content, and tag a content with an additional topic by using the three use cases *Vote A Content*, *Delete A Content*, *Tag A Content* respectively. Importantly, the *Visit A Group* use case, in addition to the *User Authentication* step, requires furthermore a *Group Access* step for ensuring that the user is one of the group's members.

5.3 User Interface

Taking into consideration the aforementioned use cases, we have created the corresponding user interface with a number of dedicated pages. We have moreover added several extra interface components for giving the user some additional useful features. We will go through all of them in this section.

5.3.1 Navigation Bar

The interface of SoCoSys is for now quite simple to facilitate the user's tasks. Its navigation bar is only composed of three main menus, namely *Home*, *Groups*, *Settings* and a *Help*
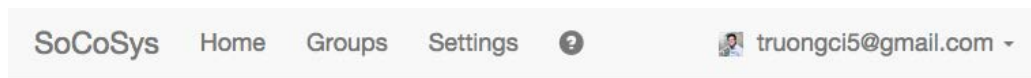
FIGURE 5.7: SoCoSys menu
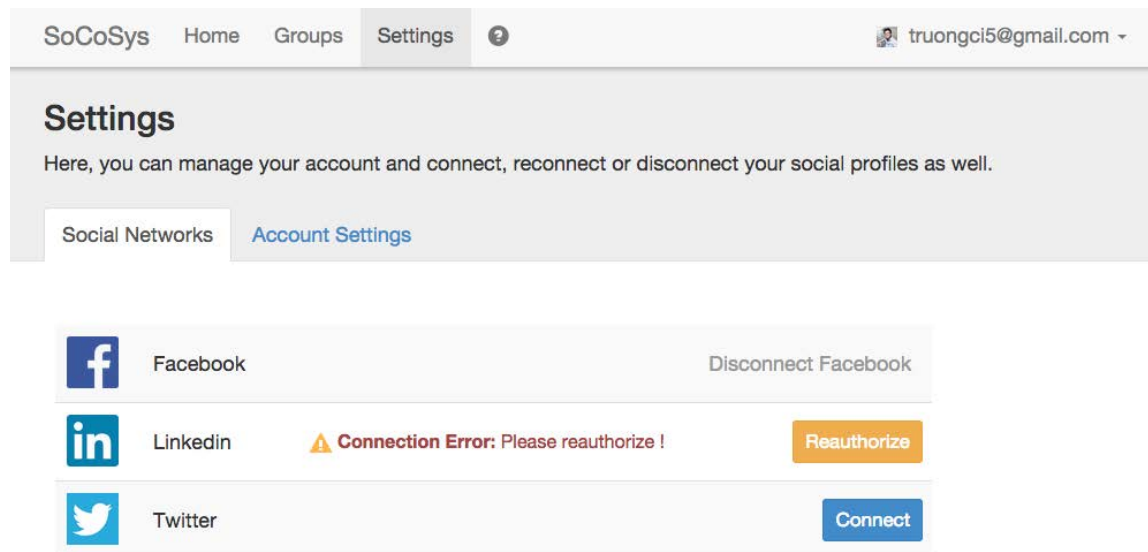
FIGURE 5.8: SoCoSys settings page

feature as shown in Figure 5.7.

Clicking on the *Home* menu leads the user to the page dedicated to the visualization of his/her aggregated social data. Clicking on the *Groups* menu sends the user to the page used for managing his/her groups. Clicking on the *Settings* menu leads to the page for managing the user's social accounts. The *Help* feature, represented by the question mark icon (❷), is aimed at giving the user some primary helps (e.g. instructions, FAQs).

5.3.2   Manage Social Accounts

Within the *Settings* page, a user is able to connect / to disconnect / to reconnect one of his three social accounts on Facebook, LinkedIn, and Twitter respectively. Let's take the example illustrated in Figure 5.8. The user in question has connected his Facebook and LinkedIn accounts, but not his Twitter account.

To connect a social account, for example a Twitter account, the user only needs to click on the associated *Connect* button. The authentication and authorization process follows the protocols imposed by the social network providers (e.g. OAuth 1.0, OAuth 2.0). Basically, it leads the user through two main steps. First, the user is sent to the SNS where the user is asked to sign into, if not yet the case, and asked to confirm that he/she wants to continue to use SoCoSys and to grant it an access to his/her social data. Then, if the user

denies the request, he/she will simply be redirected back to SoCoSys without any effect. In the opposite case, the user will be registered by the SNS as a user of SoCoSys. An *access token*[3] will also be generated and sent to SoCoSys to memorize. The user is afterwards redirected back to SoCoSys with a successful message.

The access token provided by the corresponding SNS is not endless, but has a validation time (e.g. two months). After this time, the access token is no longer valid and usable for making API calls. As a result, SoCoSys cannot continue to aggregate the user's social data from the SNS. In our example, it is the case for the user's LinkedIn access token. To make the access token valid again, the user needs to click on the associated *Reauthorize* button. The reauthorization process is simpler than the authorization process, as both the user and the application have been already registered by the SNS. In general, the user is sent to the SNS and immediately redirected back to SoCoSys without any manual intervention. The access token is indeed not changed but refreshed which means its validation time will be extended.

Finally, to stop the aggregation of social data from a previously connected social account, the user needs to click on the associated *Disconnect* link, for example *Disconnect Facebook*. Nevertheless, it does not immediately revoke (i.e. de-register) the registered access to the user's social data on the corresponding SNS. In case the user wants to reconnect the social account, the aggregation is activated again without the authentication and authorization process. Otherwise, the access will automatically be suspended by the corresponding SNS after the validation time.

5.3.3   View Aggregated Social Data

Once the user has connected at least one of its social accounts, SoCoSys starts to aggregate his/her social data from the connected social account(s). The user can therefore view the aggregated social data within the *Home* page (see Figure 5.9). For ease of reading, the social data are arranged into five different views such as *Profile*, *Friends*, *Posts*, *Following Posts*, and *Interests* (see the component 1 in Figure 5.9). The Profile view shows the profile information (e.g photo, email, first name, last name, description, location). The four other views correspond to the four types of social data (i.e. Friends, Posts, Following Posts, and Interests), in which items are displayed in a reverse chronological order with their timestamps and their origins (i.e. the original owner and the original social network). For example, the figure 5.9 shows the Following Posts view.

---

[3]An access token is an opaque string that identifies a user, an application and can be used by the application to make API calls
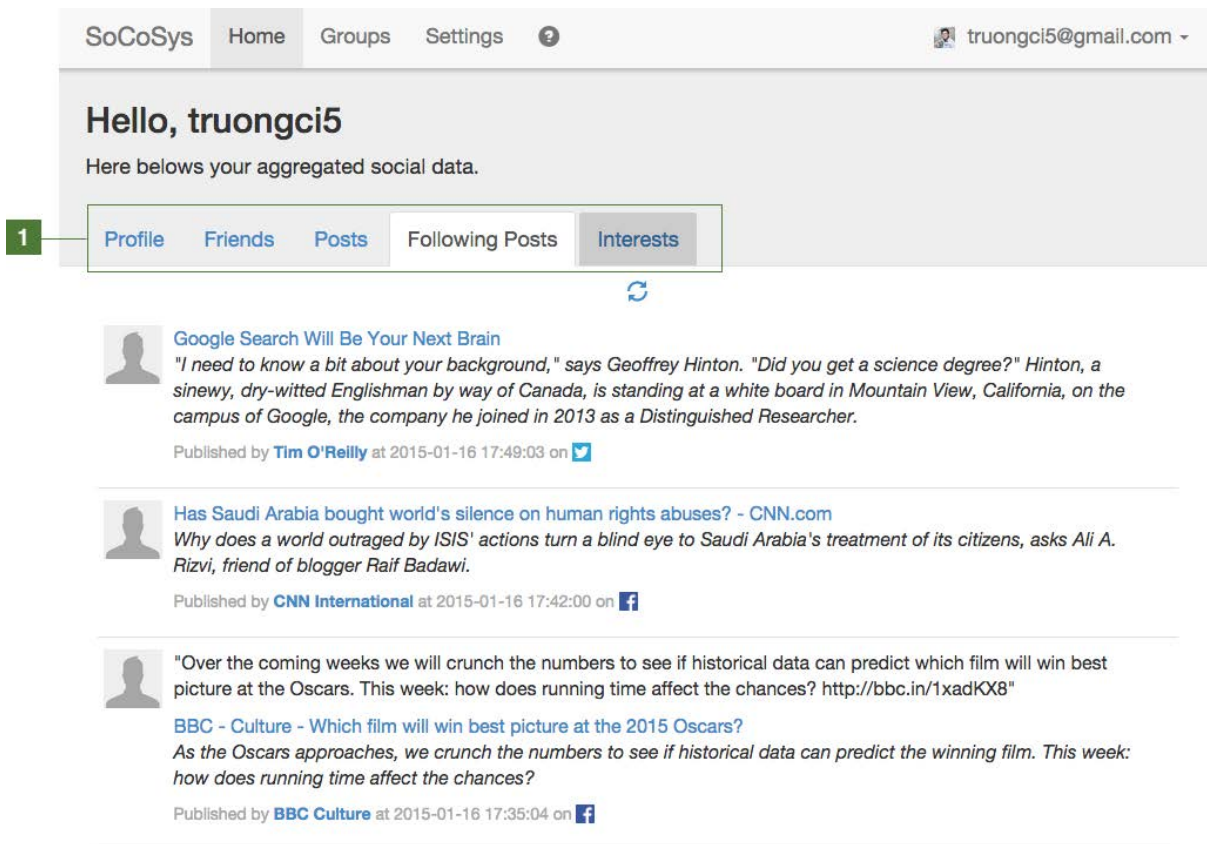
FIGURE 5.9: SoCoSys home page : (1) Different views

It is important to note that these five views are updated whenever the new social data are aggregated, and that they are exclusive to the user which means that no one else is able to see them.

5.3.4 Manage Groups

While utilizing SoCoSys, it is essential to create or to join groups in order to get and organize the contents of interest. For this task, the user needs to go to the *Groups* page (see Figure 5.10). Before creating his/her own groups, the user may begin by searching for open groups using the keyword-based search feature (see the component 2 in Figure 5.10) or by selecting one from the suggested list (see the component 4 in Figure 5.10). The search feature matches the entered keyword with the title and the description and the topics of the groups to find out the matching ones. The suggestion feature at this stage proposes randomly three open groups that the user does not yet belong to. Both features show to the user a group with its descriptive information including its name, its description, its number of members, its topics of interest, for example the Football group shown in Figure 5.11. Such information is helpful for the user to decide whether or not to join the group. To join a group, the user just needs to click on the *join* button.
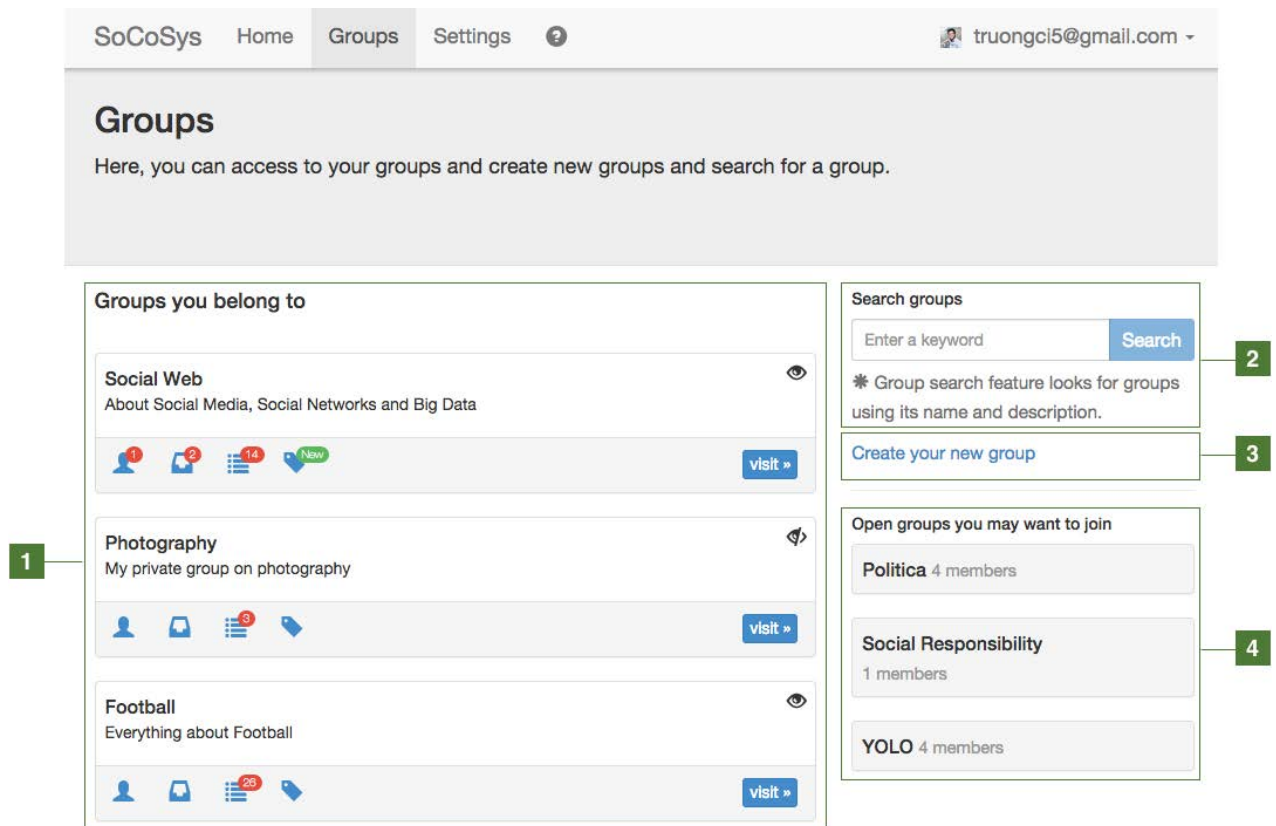
FIGURE 5.10: The groups page : (1) Groups that the user belong to , (2) Keyword-based group search feature, (3) Group creation feature, (4) Group recommendation feature
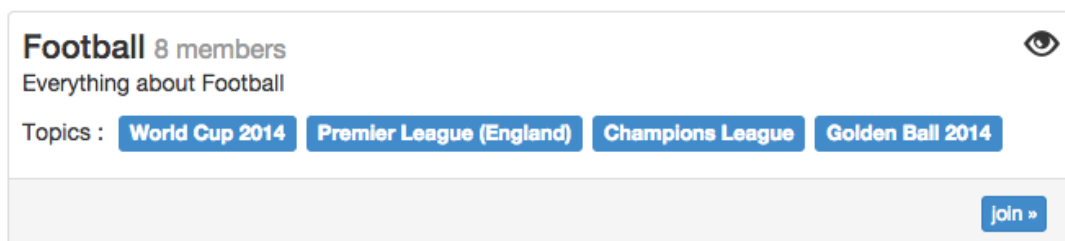
FIGURE 5.11: The group descriptive information

If the user does not find out any interesting group, he/she can create a new group by clicking on the *Create your new group* link (see the component 3 in Figure 5.10). The user will then be asked to provide a name and a short description, most importantly, to choose to make the group as private or open.

The *Groups* page also displays the list of private or open groups that the user belongs to (see the component 1 in Figure 5.10). This list is a kind of dashboard, as it gives the user an overview of the latest news and a single access point to each group. Actually, each group is associated with a number of notifications including the number of new members (👤), the number of contents to review (▭), the number of newly detected contents of interest

FIGURE 5.12: The group's notifications
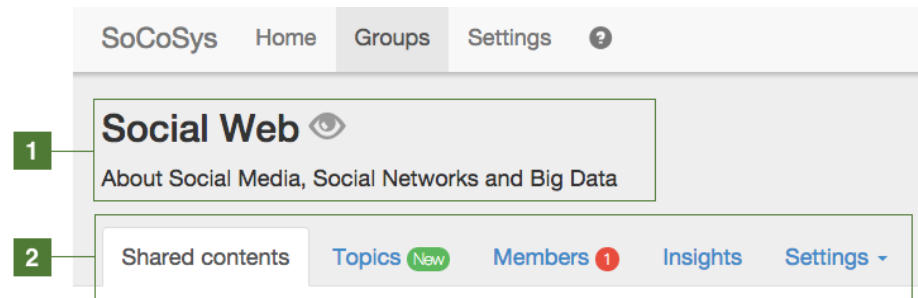
FIGURE 5.13: The group's dedicated page: (1) the group's descriptive information, (2) its different sections

(▤), and the new topics/selectors (🏷) since the last 48 hours, respectively illustrated in Figure 5.12. These notifications are good indicators for the user to decide which groups to visit first.

5.3.5   Visit A Group

Being already a member of a given group, the user can visit the space devoted to the group and participate to its management. There is no big difference in terms of interface between a private group and an open group. Both have the same number of interface components as shown in Figure 5.13. The user can visualize the descriptive information about the group such as its name, its visibility (open - 👁 vs. private - 👁‍🗨), its description (see the component 1 in Figure 5.13).

Also, the user can navigate between the group's various sections accessible under the headings *Shared contents*, *Topics*, *Members*, *Insights* and *Settings* (see the component 2 in Figure 5.13). Especially, the three first headings are possibly shown with some notifications to draw the user's attention. For example in Figure 5.13, it shows that there is one or several new topics or selectors (New) and that there is a new member (1).

5.3.5.1   Edit Sharing Settings

The first thing that the user should do after joining a new group, is to edit the default sharing settings. For that purpose, the user can use the *Edit sharing preferences* menu under the heading *Settings*. The user can modify the authorized accounts, the authorized
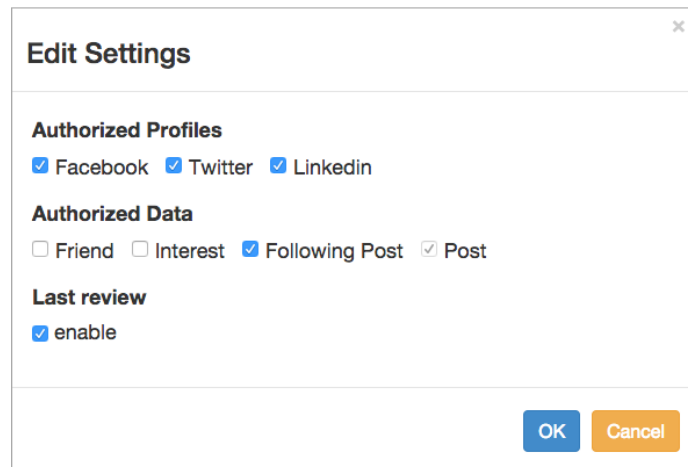
FIGURE 5.14: The form for editing the sharing settings

types of social data and enable/disable the *review* option as illustrated in 5.14. The post-type social data is activated by default, and cannot be deactivated. The user has to select at least one of the social accounts. Otherwise, the user will be prompted by an error message recalling the rule.

Also, under the heading *Settings*, the user can choose to leave the group if it no longer fits the user's needs. The user is furthermore able to either open or close the group. The former case is possible if it is a private group. The latter case is possible if, and only if, it is an open group and the user is its unique member.

5.3.5.2 Edit Topics/Selectors

The next thing to do is to add new topics and/or select interesting topics among those suggested by other members in the case of an open group. For that purpose, the user needs to select the *Topics* section, where there are two lists of topics: those that the user has followed, and those that the user has not followed (see Figure 5.15). It is possible to unfollow any topic within the first list, and to follow any topic within the second list at any time.

To create a new topic, the user has to click on the *Add another topic* link (see the component 2 in Figure 5.15). The creation form will ask the user for providing the topic name and for initializing a first selector.

As mentioned above, accepting a topic implies the default acceptance of all of its current selectors. To edit that, the user first needs to click on the topic in question, and then to unfollow the undesired selectors (Figure 5.16). When deciding either to follow or to unfollow a given selector, the user may check the information about the selector such as its creator and its recent followers displayed right by the selector (see the component 2 in
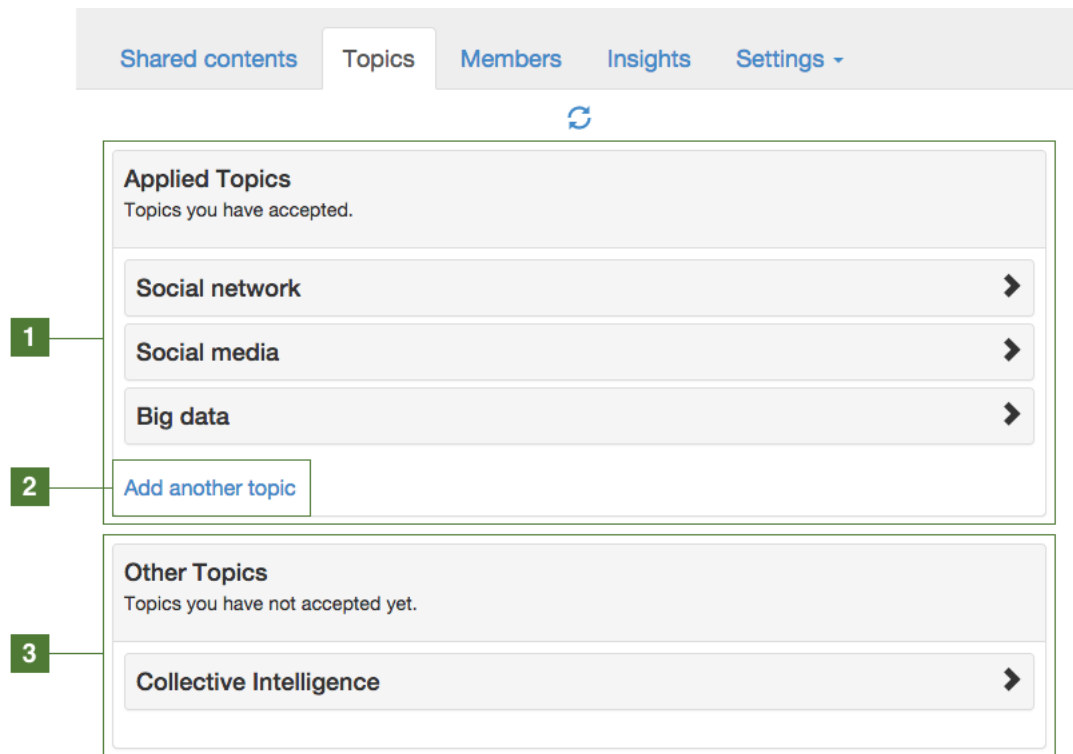
FIGURE 5.15: The group's topics: (1) the user's following topics, (2) topic creation feature, (3) other topics that the user has not followed

Figure 5.16). To add an additional selector to the topic, the user has to click on the *Suggest another selector* link (see the component 1 in Figure 5.16).

In case the user is the unique follower of a given selector (see the component 3 in Figure 5.16), he/she is also able to delete or to edit the selector (i.e. change the type and/or the value of the selector) without impacting on other members.

5.3.5.3   View Contents of Interest

The contents of interest matching the user's following selectors are shown within the *Shared Content* section (see Figure 5.17). They are displayed in a reverse chronological order. To filter the contents associated with a given topic, the user can select the topic from the list of topics shown next to the content stream (see the component 2 in Figure 5.17).

Each content of interest is shown with its containing information, often including a clickable title, an image and a short text extracted from the original web resource (see the component 1 in Figure 5.18), as well as its meta-data such as its origin, its publishing date, and its matching selector(s) (see the component 2 in Figure 5.18). All these elements are expected to offer the user a good overview of the contents.
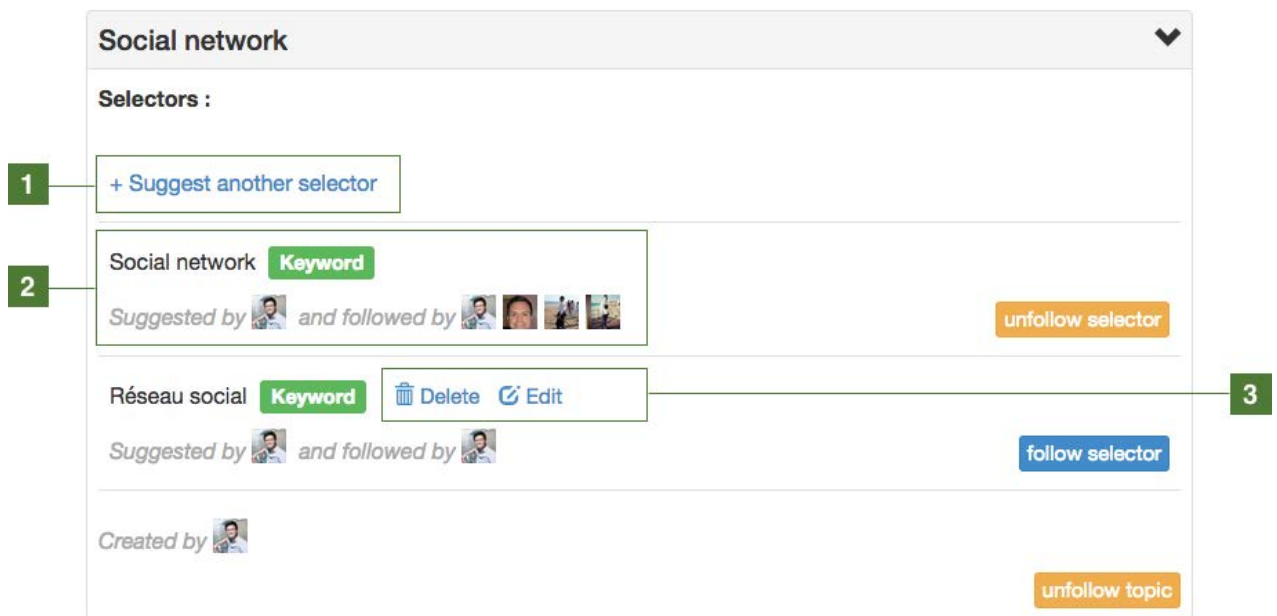
FIGURE 5.16: The selectors of a topic: (1) selector addition feature, (2) selector descriptive information, (3) selector-related features
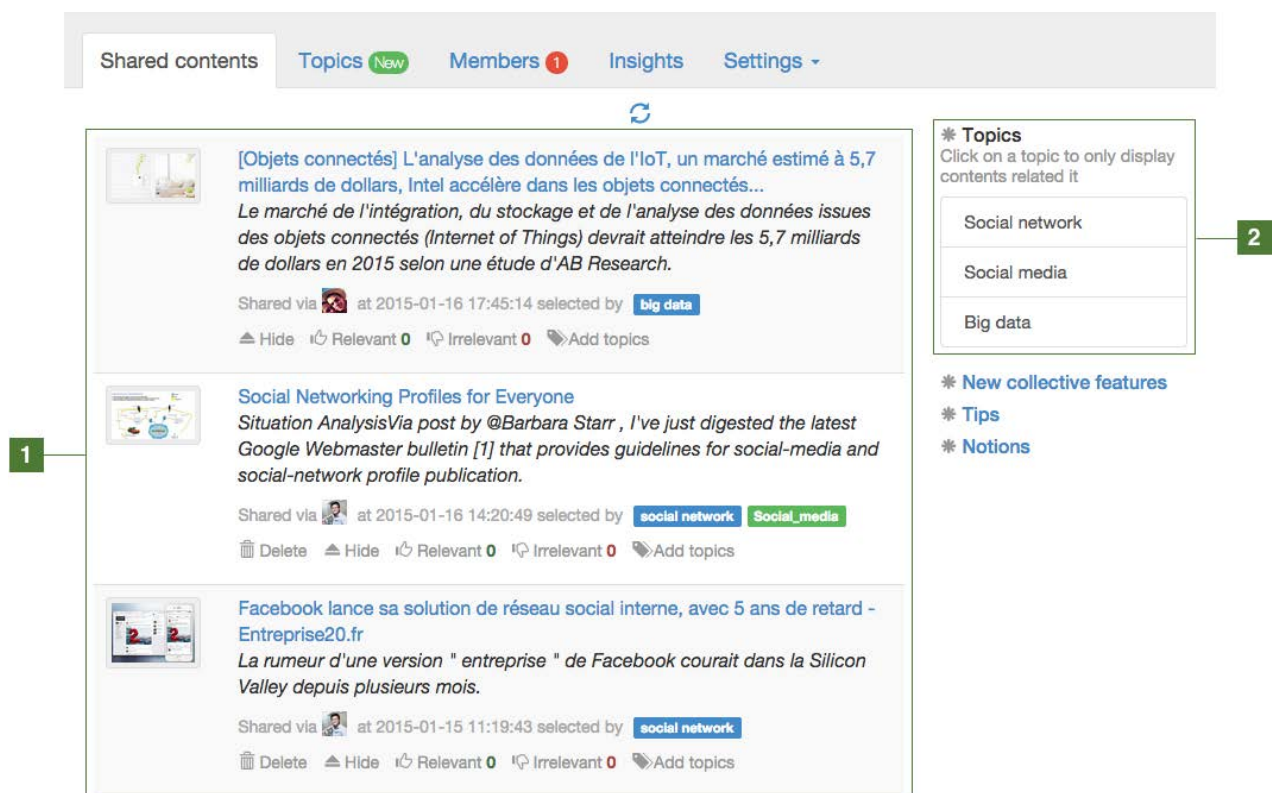
FIGURE 5.17: The group's shared contents : (1) contents of interest, (2) the user's following topics
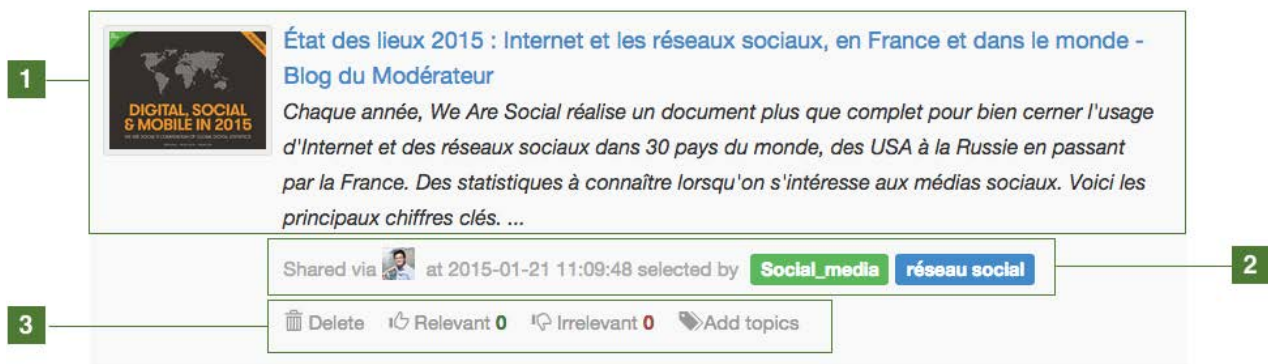
FIGURE 5.18: The content of interest: (1) containing information, (2) its characteristics, (3) collective features

Note that there is a small difference between a private group and an open group while showing the origin of a content. In the case of a private group, the original creator (i.e. the member or one of his/her social friends) of the content is shown and preceded by the expression "*shared by*". In the case of an open group, the content is not shown with the original creator, but indicated as being "*shared via*" the member, who owns it. There are two important reasons for this variance. The first reason is that the member may not choose to share with other members of the group the information about his/her social friends. The second reason is that a content being selected and shared within an open group, is under the responsibility of the member, not its original creator. It is not necessary to unveil, even important to protect the identity of its creator.

The content of interest is furthermore associated with the three permanent *collective* features including *relevant* (), *irrelevant* (), and *add topics* () buttons (see the component 3 in Figure 5.18), and eventually a *delete* button () in case the user is the owner of the content. These four features allow the user to promote or demote the content, to assign additional topics to it, and to remove it, respectively. The numbers right next to the *relevant* and *irrelevant* buttons will give the user an additional indicator to decide whether or not visit the original web resource for more information.

In case the user has enabled the *review* option, SoCoSys will keep the contents extracted from his/her social data and matching his/her following selectors for manual approval. This way, when visiting a group, the user will be prompted to review these contents, if any. The user can for each content accept or delete it (see Figure 5.19). As such, the content will be shared with the group or definitively removed.

On the other hand, it is possible that a big number of contents of interest have been detected during the time between the user's two successive sessions (e.g. a week or longer). The user may then feel frustrated by spending a lot of time to review these contents before being able to view the group's shared contents of interest. To reduce the user's review time, we
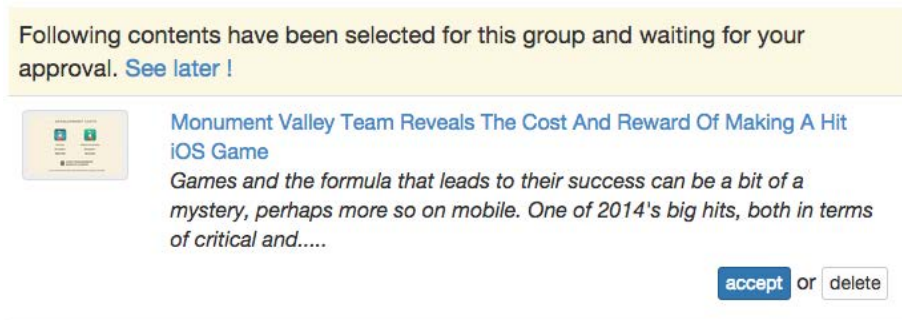
FIGURE 5.19: A content to review

chose to remove the waiting (suspended) contents dated more than 4 days. Such choice is based on two assumptions: first, these contents may become outdated already, thus less interesting; second, even if they are validated and shared with the group, given their old created date, they are probably swamped by other more recent contents, thus receiving less attention from the other members.

5.3.5.4 View Members

In addition to the interface components derived from the predefined use cases, we have also included several additional features, one of which is *viewing members* accessible under the heading *Members*. This feature allows the user to access to the entire list of members of a given group. For each member, it shows the topics of interest that he/she has chosen, associated with the number of matching contents extracted from his/her social data for the last two weeks (see Figure 5.20). Thereby, we can have a rough idea about the degree of involvement and participation of each member. We can, for example, know who are the *active* contributors, who are the *passive* consumers.

5.3.5.5 Get Insights

The second additional feature is accessible under the heading *Insights*. It is aimed at giving the group some significant insights of its sharing activities. At the moment, we have simply applied some basic statistics and displayed them as graphics.

The user can visualize three different charts. The first chart called *topic evolution* shows the evolution in volume (i.e. quantity) of every topic of the group over the last 30 days (see Figure 5.21). The second chart called *topic repartition* shows the repartition in percentage of each topic of the group (see Figure 5.22). The third chart is a table showing the top three trending topics and their best contributors (see Figure 5.23).
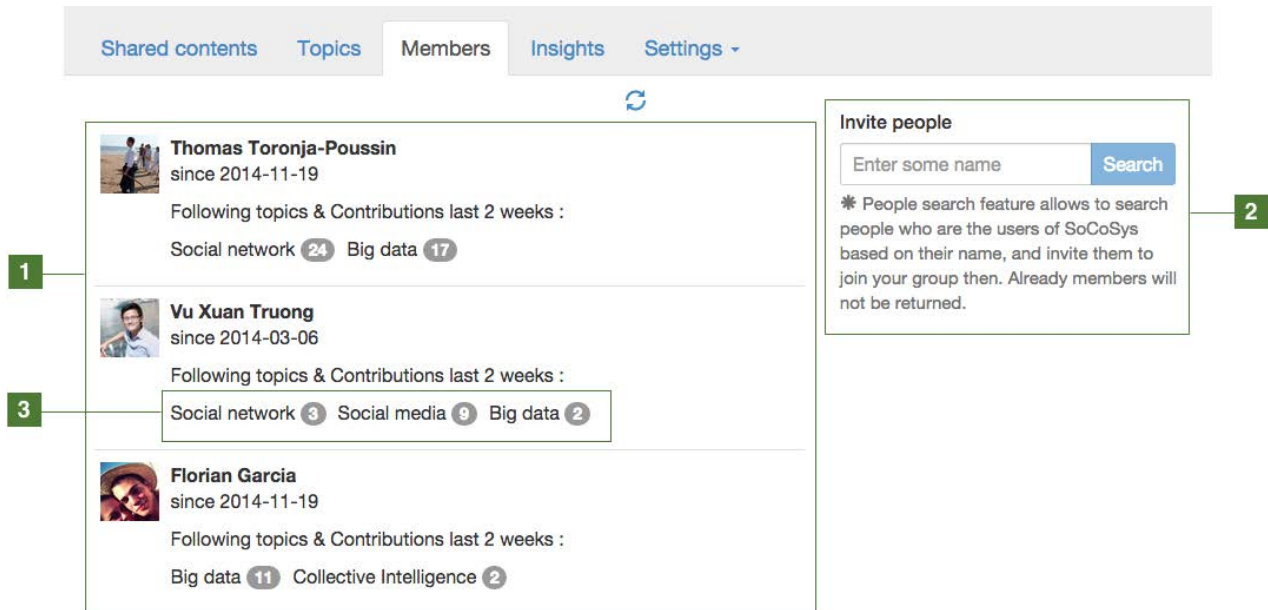
FIGURE 5.20: The group's members: (1) current members, (2) invitation feature, (3) member's recent contribution counting

Using these three graphical elements, we can estimate which topics are the main interests of a group, and which members are the experts of such topics. Let's consider the example shown in Figures 5.21, 5.22, and 5.23, we can say that the group in question potentially has good expertise, competency and resources in the areas of *Big Data* and of *Social Network*.

5.4   Summary

We have shown, in this chapter, a first prototype, called SoCoSys, of our proposed solution to the three addressed questions. It has been deployed as a Web-based application with responsive interfaces so that the users can access and utilize it anywhere and on different devices. Despite its simplicity due to the objective of facilitating the user's tasks, the user Web interface of SoCoSys derived from the predefined use cases, have fully complied with the conceptual requirements.

Actually, this user Web interface adopts a simple style of web navigation which is the navigation bar. The three main menus shown in the navigation bar, namely *Home*, *Groups* and *Settings* give access to three different pages allowing the user to view his/her aggregated social data arranged into five various views (i.e. profile, friends, posts, following posts, and interests), to mange his/her social accounts (i.e. Facebook, Twitter, and LinkedIn accounts), and to manage his/her different groups, respectively. The Groups page moreover provides the user with a single access point and the notifications of the latest activities of each of his/her groups. The page dedicated to a given group is also organized with the

Topic evolution

This line chart shows the evolution in volume of every topic of the group over the last 30 days.
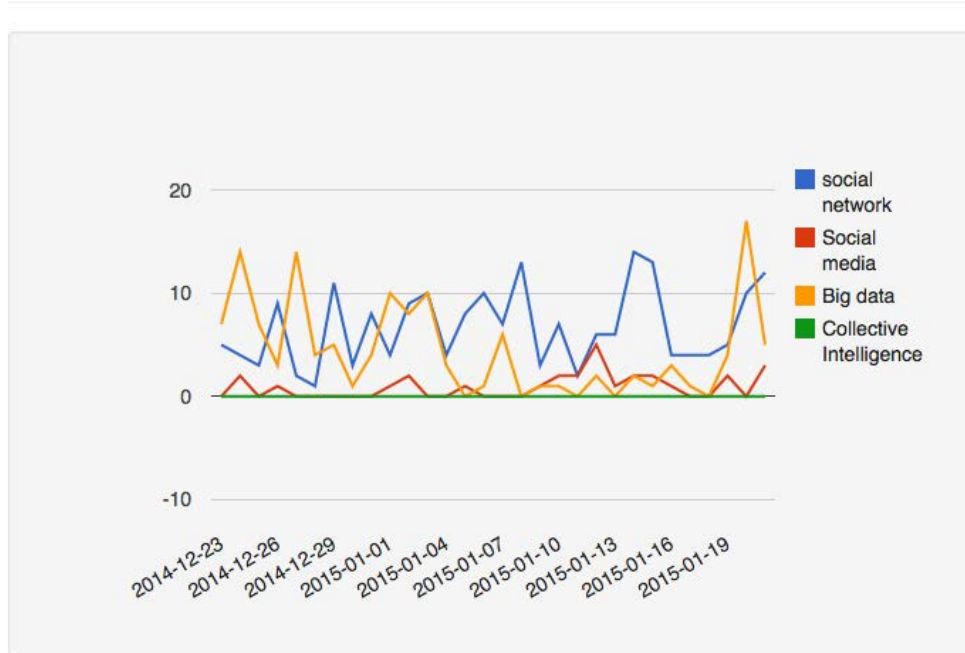
FIGURE 5.21: The topic evolution chart

Topic repartition

This pie chart shows the repartition in percentage of each topic of the group.
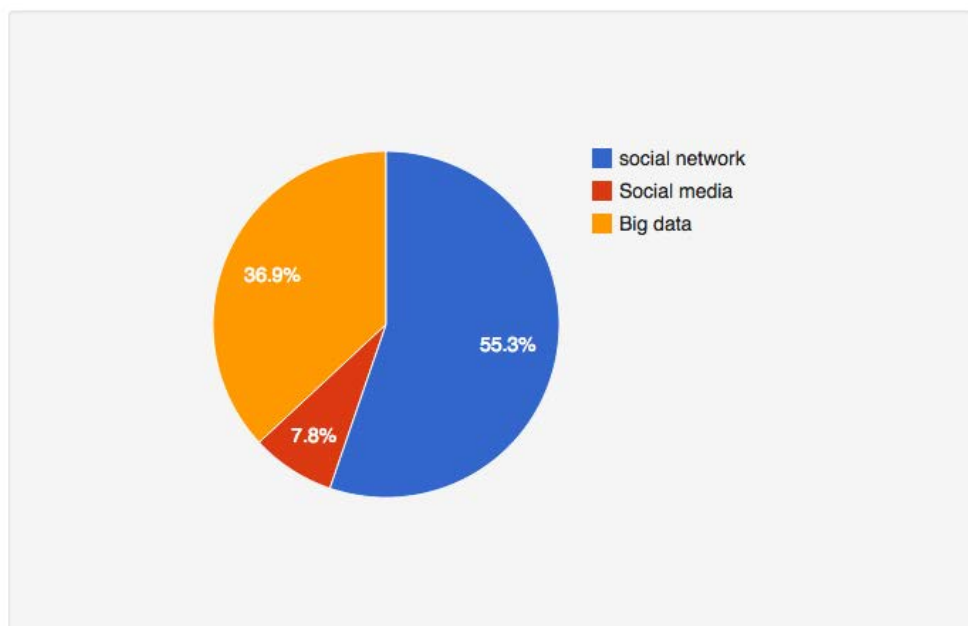
FIGURE 5.22: The topic repartition chart

Trending topics

This graphic shows some trending topic and members sharing the most about them as well.

Big data

Social network

Collective Intelligence
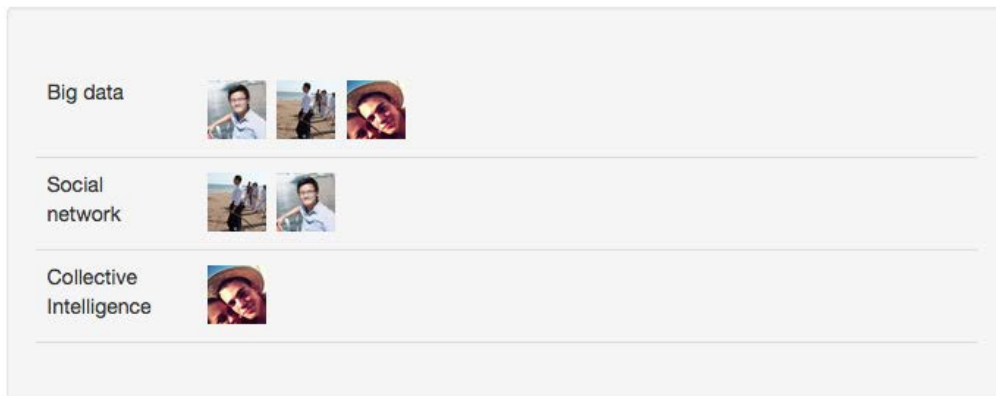
FIGURE 5.23: The trending topics

same navigation bar principle. The member can edit his/her sharing settings, add and/or follow/unfollow the different topics and their associated selectors, and view and react to the group's shared contents of interest from the different sections, namely *Settings*, *Topics*, and *Shared contents*.

In addition, we have also added two new group-specific features which are viewing members and getting insights. Although, both features are not directly linked to the sharing purpose, they are useful for the group. They provide the group with primary means for accessing to the advanced knowledge on its internal collaboration such as its evolving centers of interest, and the individual expertises of its members.

For the next chapter, we will present an experimental evaluation of our prototype by means of two small tests with two different sets of users.

Chapter 6

Experimental Evaluation

Contents

To evaluate our proposed approach and the developed system, we have carried out two small tests with two different groups of users. Although, both tests were realized using the same web-based prototype SoCoSys, they have different purposes. In this chapter, we will detail these two tests, their settings as well as their results. Then, we will discuss about some interesting suggestions derived from the two tests, and also their limitations.

6.1   First Test

6.1.1   Settings

The first test has a twofold purpose: to serve as a functional test to detect and fix possible errors, bugs, ambiguous points; and to provide real data, based on which we can quantitatively analyse the use of popular social networks and their wealth of information.

The test group consisted of ten volunteered participants ($n = 10$). Most of them are international PhD students at the University of Technology of Compiègne. They were selected as they are regular users of social network sites. We introduced the dedicated interfaces and the main operation of SoCoSys to each of them. We furthermore assisted them during the test for understanding and performing various features.

6.1.2   Statistical Analysis

During one month of testing, from June 1st to 30th 2014, we could identify several issues that we fixed as early as possible. After the testing period, we obtained a set of real data, based on which we made a number of representative statistics.

In Table 6.1, we present some important figures related to the connected social accounts.

TABLE 6.1: Statistics on social accounts

Indicator	Total
Number of connected social accounts	19 ($\approx$2 per person)
Number of Facebook accounts	10
Number of Twitter accounts	4
Number of LinkedIn accounts	5
Number of participants with 3 connected accounts	2
Number of participants with 2 connected accounts	4
Number of participants with 1 connected account	4

We had 19 connected social accounts for 10 participants, almost 2 per person. Especially, all participants granted SoCoSys an access to their Facebook accounts. It may be understandable that they all consider Facebook as an important and principal source of information and contents of interest. Six out of ten participants were also connected with one another profile, Twitter or LinkedIn or both, which shows their interests of aggregating their different social networks.

Table 6.2 shows the averaged numbers about the participants' social data (i.e. friends, posts, and following posts) aggregated for one month from the three social networks Facebook, Twitter and LinkedIn. The numbers related to Facebook confirm again its importance

both for networking and communicating. On average, a participant had 300 friends on Facebook and received nearly 100 pieces of following posts per day. The participants had less friends on Twitter than on Facebook but tended to receive more contents from their following friends. This can be explained by the fact that Twitter plays an increasing role in publishing and sharing information and contents. The participants had very little activity (i.e. posting and receiving) on LinkedIn while having a significant number of connections, probably due to the business oriented characteristics of LinkedIn.

TABLE 6.2: Statistics on social data

Indicator	Average (per person)
Number of Facebook Friends	300
Number of Twitter Friends	120
Number of LinkedIn Friends	140
Number of Facebook Posts	3
Number of Twitter Posts	16
Number of LinkedIn Posts	1
Number of Facebook Following Posts	3000 (100 per day)
Number of Twitter Following Posts	3450 (115 per day)
Number of LinkedIn Following Posts	50 (2 per day)

An average participant received nearly 100 posts from his/her Facebook friends and 115 posts from his/her Twitter friends per day. Such quite big numbers normally require the participant to spend considerable time and effort to manually select the interesting contents. Some participants could therefore be overwhelmed, thus ignoring lot of incoming information.

On the other hand, we found with surprise that about 90% of following posts contained at least one URL. Such a very high percentage confirms that the social networks like Facebook or Twitter represent a powerful source of information which needs to be efficiently exploited.

TABLE 6.3: Statistics on groups

Indicator	Value
Number of groups	10
Number of private groups	6
Number of open groups	4
Number of participants joined at least one open group	8
Average number of members of an open group	4

The participants created 10 groups in total, 6 private groups and 4 open groups (Table 6.3). Eight out of ten participants joined at least one open group. Most of them did not edit their respective sharing settings and also disabled the *review* option, which could be explained by the fact that they already knew and trusted each other.

On average, an open group had 4 members (i.e. 40% of the participants). The two most "successful" groups were the groups of *Football* and *Politics* which gathered 6 and 4 members, respectively. This is totally understandable given the very broad and common centers of interest of the two groups.

TABLE 6.4: Statistics on topics

Indicator	Value
Number of topics	25
Average number of topics per group	2.5
Average number of topics per private group	2
Average number of topics per open group	3.5
Number of selectors	64
Average number of selectors per topic	2.5
Average number of selectors per topic in private groups	1.5
Average number of selectors per topic in open groups	3.5
Number of keyword-based selectors	48
Number of hashtag-based selectors	15
Number of concept-based selectors	1

Table 6.4 shows some numbers on the topics of interest of the participants. The participants, in total, created 25 topics (i.e. 2.5 per group), and created 64 selectors (i.e. 2.5 per topic). These numbers quantitatively shows the participants' wide range of interests.

With no surprise, there were more topics in open groups (i.e. 3.5 topics/group) than in private groups (i.e. 2 topics/group). The topics of open groups were also associated with more selectors than those of private groups. This shows the collective efforts within the open groups.

In addition to the aforementioned statistical analysis, we also took a deeper analysis on the scope of the added topics and the behaviours of the participants when creating topics and selectors. We found that the topics of interest varied a lot from general areas like *Football*, *Politics* to specialized areas like *Social Media* and *Social Responsibility*. More specially, in 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".

Regarding the selectors, the participants mostly used keyword-based selectors and hashtag-based selectors. This is probably because they are already familiar with these techniques. Especially, the participants did not accept systematically every selector but selected well those corresponding to their interests.

We also observed that there were two strategics when suggesting selectors: adding multi-language or synonymous terms, or 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, we counted the number of contents of interest that each group received during the testing period. We saw that all groups, no matter what their centers of interests are, received a number of contents of interest extracted from their collective sources of social data. The number varied from 34 to more than 300 pieces of contents according to the topics of interest. This approves again that social network sites like Facebook or Twitter represent an important and multi-domain wealth of information that needs to be efficiently exploited.

During the test, with the help of the participants, we detected a number of divers bugs and errors related to the different components of the application (e.g. user interface, web services, back-end system). These reported issues were solved as soon as possible to reduce the impact on the participants' user experience. At the end of the test, no major issues were reported, and SoCoSys was completely operational.

6.2 Second Test

6.2.1 Settings

The second test was carried out in order to supplement the findings obtained from the first test about the use of popular social networks. It furthermore aimed at evaluating the proposed system (SoCoSys) in terms of *utility*, *functionality* and *usability*.

To avoid bias, we decided not to solicit the same participants of the first test again, but to invite new users. We therefore chose a group of third-year engineering apprentices[1] at the University of Technology of Compiègne. Unlike regular students, they also follow a part-time professional training in companies, thus having probably a more practical and functional point of view. Such a characteristic is very interesting for our test.

We invited 13 engineering apprentices, aged between 22 and 26, to a presentation of SoCoSys followed by a detailed demonstration. Two of them claimed not to have any account on Facebook, Twitter, or LinkedIn, thus could not test the system. Out of the 11 sent invitations, 7 (64%) students started to test SoCoSys (the same version of SoCoSys at the end of the first test). After 3 weeks of use, from November 20th to December 11th 2014, we sent questionnaires to these 7 students for obtaining their feedbacks that we will detail in the next subsection.

---

[1]in French "apprentis ingénieurs"

6.2.2   Questionnaires

The complete version of our questionnaires is in French and available in Appendix French Questionnaires. They are composed of two main parts: (1) the use of social networks, and (2) the use of SoCoSys.

6.2.2.1   Use of Social Networks

The first part contains six questions as shown in Table 6.5. For the first question, all the participants answered that they used all of the three social networks, namely Facebook, Twitter, and LinkedIn. This proves again the popularity of these three social networks.

TABLE 6.5: Questions on the user of social networks

Q1	Do you have at least one profile on the following social networks ?		
	Facebook	7	100%
	Twitter	7	100%
	LinkedIn	7	100%
Q2	If you use multiple social networks, what is your habit ?		
	You try to visit all as frequently as possible	0	0%
	You mainly use a given social network and visit occasionally others	7	100%
Q3	What is your current frequency of using social networks ?		
	Many times a day	3	43%
	At least one time a day	3	43%
	Several times a week	1	14%
	From time to time	0	0%
Q4	Do you share information and contents of interest on your social networks ?		
	Not at all	0	0%
	A bit	6	86%
	Much	1	14%
Q5	Do you think that your social networks bring you a lot of interesting information ?		
	Not at all	1	14%
	A bit	5	71%
	Much	1	14%
Q6	Do you think that there is an information overload on your social networks ?		
	Yes	5	71%
	No	2	29%

For the second question, the participants, once again, gave the same answer which states that they mainly use a given social network and visit occasionally others. This may be explained by the fact that using simultaneously different social networks is not really convenient and requires much more time. This also means that the participants obtain many interesting information from their principal social network, but probably ignore other interesting information published on other social networks.

For the third question, we had 3 responses for *many times a day*, 3 responses for *once a day*, and 1 response for *several times a week*. So, more than half of the participants do not want or can not spend so much time on social networks. As a result, between their two consecutive visits, certain interesting information is probably ignored.

For simplicity, the questions 4 and 5 were only given with three possible choices, *not at all*, *a bit*, and *much*. All the participants have shared information and contents of interest on their social networks (e.g. 6 *a bit* and 1 *much*). Most of them (6/7) agreed that they could get interesting information from their social networks (e.g 5 *a bit*, 1 *much*). Based on these numbers, we can see that the users started to consider their social networks as an important source of information and contents of interest.

Note that we stated these two questions in a general sense, and did not deepen the two questions further regarding the types of contents published on social network sites. The participants were furthermore asked to answer the two questions based on their personal and overall impression and satisfaction. Otherwise, given the multifaceted nature of the published contents and the users' different interests and personal satisfaction degrees, a complete study is first necessary to exclusively and objectively categorize the published contents.

For the last question of the first part, the majority of the participants (i.e. 5/7) believed that there is an information overload on their social network.

To sum up, the responses of the participants to this first part of questionnaires totally comply with our initial assumptions on social networks, their potential sources of information and their problems of multiple walled networks and information overload.

6.2.2.2 Use of SoCoSys

The second part of questionnaires includes in total 14 questions as shown in Tables 6.6, 6.7.

For the first two questions, the participants had three possibilities, *yes*, or *may be*, or *no*. No one answered negatively to these two questions. Nevertheless, they chose *yes* and *may be* answers in a quite equal manner, for instance, it was 3-4 for the first question and 4-3 for the second question. The first tight score can be explained by two facts. The first fact is that the participants are used to visiting, and obtaining the information from one principal social network while ignoring others. The second fact is that they may be worried about the privacy and the security of their whole social data stored within a single place. The second tight score may be linked to the time constraint. The participants may need more time to use SoCoSys, especially to collaborate within some open and interesting groups.

TABLE 6.6: Questions on the use of SoCoSys (I)

Q1	Would you find useful to aggregate your social networks, to extract the interesting information and to make it accessible at a single location ?		
	Yes	3	43%
	No	0	0%
	May be	4	57%
Q2	Would you find interesting to share information extracted from your social networks with your groups of interest ?		
	Yes	4	57%
	No	0	0%
	May be	3	43%
Q3	Do you think that SoCoSys offers both the aforementioned features ?		
	Yes	6	86%
	No	0	0%
	May be	1	14%
Q4	Is it generally easy to use the Web interfaces of SoCoSys ?		
	Yes	6	86%
	No	1	14%
Q5	Is it generally simple to understand how SoCoSys works ?		
	Yes	7	100%
	No	0	0%
Q6	Do you agree with the organization by groups (private versus open) ?		
	Yes	7	100%
	No	0	0%
Q7	Do you think that the current filtering mechanism is good ?		
	Yes	6	86%
	No (more automation)	1	14%

Most of the participants (6/7) totally agreed that SoCoSys offered the two possibilities for extracting contents of interest from their social networks, and for sharing them within groups of interest. The last one did not say the opposite, and thought that it may be the case.

If the three first questions recall the participants the utilities of SoCoSys, the two questions 4 and 5 are about its usability. All the participants thought that it was generally easy to use the Web interfaces of SoCoSys and most of them (6/7) thought that it was generally simple to understand how SoCoSys works.

The six following questions are aimed at asking the participants for personal opinions on certain conceptual points and technical choices of SoCoSys. They totally agreed with the organization by groups (private versus open), the ability to limit what can be shared in an open group, the collective definition of topics of interest within an open group, and the ability to personalize the topics of interest in an open group.

Six participants thought that the current filtering mechanism, in which the user manually adds his/her topics of interest as input and the system completes the rest, is relevant. Only

TABLE 6.7: Questions on the use of SoCoSys (II)

Q8	Do you agree with the ability to limit what can be shared in an open group ?		
	Yes	7	100%
	No (every one should open every thing)	0	0%
Q9	Do you agree with the collective definition of topics of interest within an open group ?		
	Yes	7	100%
	No (only qualified members)	0	0%
Q10	Do you agree with the ability to personalize the topics of interest in an open group ?		
	Yes	7	100%
	No	0	0%
Q11	In your opinion, the three current filtering methods (i.e. keyword, hashtag, and concept) are enough ?		
	Yes	6	86%
	No	1	14%
Q12	In your opinion, should SoCoSys be rather for personal use or collective use or both ?		
	Personal use	1	14%
	Collective use	2	29%
	Both	4	57%
Q13	Do you think that SoCoSys can also be deployed in organizations/companies as a collaborative working tool ?		
	Yes	6	86%
	No	1	14%
Q14	Finally, do you want to continue using SoCoSys after the test ? If not, why ?		
	Yes	2	29%
	No	2	29%
	May be	3	43%

one said that it should be automated completely. Likewise, 6 participants though that the three current filtering methods (i.e. keyword, hashtag, and concept) were enough. Only one said that additional methods should be included.

The questions 12 and 13 deal with the scope of the use of SoCoSys. More than half of the participants (4/7) believed that SoCoSys could be used for both personal and collective purpose. Moreover, most of them (6/7) thought that SoCoSys could be also be deployed in organizations/companies as a collaborative working tool, for example for a collaborative technological watch.

For the last question, we obtained from the participants 5 responses for *yes* and *may be* that they keep using SoCoSys after the test. There were two negative answers but with interesting explications. The first one said that SoCoSys till needs to be ergonomically improved. The second one argued that his/her social networks are too small to really see the benefits of SoCoSys.

In short, the general opinions of the participants toward SoCoSys are very positive. The participants mostly agreed with the utilities of SoCoSys, its conceptual and technical points, and its dedicated interfaces.

6.3   Suggestions

Throughout the two tests, we have also received from the participants a number of interesting suggestions which focus on various aspects. These suggestions have led to certain additional functionalities worth further consideration. Here, we will take a look at some of them.

6.3.1   Protected Groups

The choice between the two current types of group (i.e. private and open) seem to be limited in some cases. Some users may need to create a group, in which they can collaborate while restricting its access to unexpected people. Both private and open groups do not meet such requirement. A third type of group, *protected groups*, should be considered. Like open groups, protected groups are not hidden, but visible for other users. However, to join a protected group, a user would have either to receive an invitation from one of its current members, or to submit a membership request which should be approved by one or several members of the group.

6.3.2   Group Recommendation

At present, the current group suggestion feature randomly shows a small number of open groups, of which the user is not yet a member. In the opinion of some participants, when the number of open groups grows, this feature should be more personalized. It should recommend open groups, even protected groups suitable for the user. For that purpose, it could rely on the topics of interest of the user added in his/her different groups and the topics of interest of the group.

6.3.3   Duplicated Information

During the two tests, we have repeatedly observed that many retrieved contents, especially in open groups, are redundant. These contents are extracted from the social data of different members who are shared by different sources, but refer to the same information. Even though, it means that the information is important, the users are probably bored

by repetitively viewing similar contents. Thus, we should investigate means to highlight the important information while hiding the duplicated contents. In the short term, we can include a new collaborative feature, which is a *hide* button associated to each content of interest right to the *relevant*, *irrelevant*, *add topics* buttons. Any member of the group can use this feature to report a repeated content, thus hiding it.

6.3.4 Reporting A Source

In addition to the duplicated information issue, we have also discovered another less frequent issue but worthy of consideration. By manually inspecting a given sample of voted irrelevant contents, we found that a big number of false positives (contents retrieved by the retrieval module which are not really interesting) were published by several specific sources (social network users that the members befriend or follow). If such sources can be detected and discarded over time, the number of irrelevant contents may be considerably reduced. For that, a possible solution is to include another new collaborative feature, which is a *source reporting* button. Using this button, the members can report a source as a bad source. A source should be definitively discarded from the group's collective social data sources after a certain number of reports.

6.3.5 Sharing Back

When visiting a group, in particular an open group, a user may find already viewed contents, but most importantly, discover new contents. The user may find a given content particularly interesting and feel the need to share it with, for example, his Facebook friends without having to visit Facebook. This is not possible with the current version of SoCoSys, but is not hard to implement. The only thing to take into consideration is to prevent the contents, which have previously been shared, from being selected once again by SoCoSys.

6.3.6 Notifications

The users are busy, and thus cannot regularly visit SoCoSys. To help the users to stay current with interesting information, it could be convenient to include an email notification feature. With a personalized frequency, it notifies the users of the new activities (e.g the newly detected contents, the newly added topics, etc.), if any. Moreover, when the user visits SoCoSys, the notifications should be more explicitly and actively pushed to the user to draw his/her attention to certain groups and/or certain contents of interest, obviously with personalized settings possibilities.

6.4   Limitations

These two small tests allowed us to evaluate our proposed approach and the developed system on many aspects, and provided encouraging results. Nevertheless, they have some limitations. Actually, we have not been able to properly evaluate the two important criteria, *effectiveness* and *scalability*.

6.4.1   Effectiveness

In our case, evaluating the effectiveness of the system mainly means evaluating the performance of the filtering process, which has been, as mentioned above, implemented using the Information Retrieval techniques. Many different measures for evaluating the performance of information retrieval systems have been proposed. *Precision* and *Recall* are the two most common and important measures [122]. *Precision* is the fraction of the retrieved documents relevant to the user's information need (see Equation 6.1), whereas *Recall* is the fraction of the documents relevant to the query that are successfully retrieved (see Equation 6.2).

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{6.1}$$

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \tag{6.2}$$

Both measures require the knowledge of all stored documents to compute the number of *relevant* documents. It is not trivial when it comes to investigating the social data, which are constantly appearing in large numbers. Even if we are able to do that, we should first obtain the permissions of the users for reading their entire social data, some of which may contain sensible information. Moreover, the notion "relevant" is dynamic, and depends on the person in question and his/her current context. A piece of content may be found relevant by one but irrelevant by another according to their respective need and expertise degrees.

Therefore, in our case, it would be more interesting to measure the *personalized relevancy* of contents of interest. It is the percentage of the contents that are, in the user's opinion, relevant to a given topic, on the total retrieved contents.

It is actually possible to do that with the current version of SoCoSys. The user can use the *delete*, *irrelevant* buttons to indicate the irrelevant contents. Given the number of irrelevant contents, it is easy to compute the percentage of the relevant contents, and subsequently the *personalized relevancy* of contents of interest for each user. Because of lack of time, the

participants of the first and second tests ignored to use these features for the evaluation purpose. They nevertheless confirmed that there were false positives among the retrieved documents, but in a very small ratio.

6.4.2 Scalability

At the moment, SoCoSys supports only three social networks (i.e. Facebook, Twitter, and LinkedIn), and aggregates and processes solely the social data corresponding to the predefined common model. It has furthermore been only used by a small number of users (i.e. less than 10 concurrent users). Thus, SoCoSys has until now worked correctly without remarkable response delay time.

However, we cannot make sure that this correct response time would be ensured when the number of users increase. It is important to investigate how scalable SoCoSys is for handling the increasing demand. In our case, to measure the system scalability, the following subsequent criteria should be considered:

- The execution time of the aggregation task (including the enrichment and the indexing steps) according to the number of connected social accounts, the number of users, and the number of supported social networks,

- The execution time of the searching task according to the number of selectors, the number of topics, and the number of groups,

- The maximum number of concurrent connections,

- The response time of various actions performed on the user Web interface.

To this end, we obviously need more users and data.

6.5 Summary

In order to evaluate our proposed approach and the developed system (SoCoSys), we have carried out two different tests with two different test groups. The two tests provided us with many encouraging results which allow us to confirm our research assumptions. The data obtained from the first test and the participants' responses to the first part of the questionnaires of the second test confirm that the two addressed problems are real, and show that social networks are very potential sources of information and contents of multiple domains. In addition, the participants' positive opinions on the use of SoCoSys approve its utilities (i.e. filtering and sharing), its usability, and its functionalities.

Because of their small sizes and short testing times, we have not been able to properly measure the effectiveness and the scalability of SoCoSys. However, the generally good feedbacks of the participants of the two tests indicate that it works correctly for the present time.

Thanks to the two tests, we have also obtained a number of interesting suggestions. Some of them lead to certain additional features that need to be studied further and be eventually included in the next versions of SoCoSys.

Chapter 7

Perspectives and Future Work

Contents

It is important to recall once again that our user-centered and group-based approach for social data filtering and sharing is novel and unique. The developed system, SoCoSys, is, at this stage of the project, a proof of concept of the proposed approach. Thus, different interesting perspectives are possible and worthy of consideration for future work. In this chapter, we mention some of these perspectives grouping them in two groups: short-term perspectives and long-term perspectives.

7.1 Short-term Perspectives

The current version of SoCoSys has been designed as a proof of concept demonstrating the benefits of our proposed approach. Thereby, it has been implemented with a number of specified modules, some of which are for now employing generic and simplified techniques, thus being improvable. Although, the experimentation with two different test groups, presented in the preceding chapter, has shown promising results, there is a need for the improvement of the current version.

In the short term, we will study further, in terms of benefits and feasibility, the suggestions provided by the participants of the two tests. Some of them can then be included in the next versions of SoCoSys, which will be tested with bigger groups of users.

On the one hand, these experiments will provide us with new sets of data (i.e. raw data and participant feedbacks) large enough for a complete evaluation on the system effectiveness and scalability, which are currently the limitations of this work. Also, they will allow us to identify some potential technical drawbacks and performance gaps of the system. The corresponding improvements, for example those mentioned in the chapter *Technical Solution*, could then be applied.

On the other hand, we would like to carry out at least one test within an organization or an enterprise. Such a test is aimed at exploring the possibility to use SoCoSys as a collaborative tool for extending the internal collaboration of the organization to some open and popular SNSs, and especially for collaborative technological survey.

7.2   Long-term Perspectives

While our short-term perspectives focus on the extended evaluation and the improvement of the currently implemented system, our long-term perspectives will address some more fundamental aspects of the proposed approach. Indeed, we envision two major directions: the first direction attempts to extend the initial scope of the approach (i.e. filtering and sharing) to the *group-specific knowledge discovery*; the second direction examines the possibility of transforming the system architecture from the current centralized configuration to a distributed configuration in order to make the approach more scalable, interoperable. In this section, we will present the respective underlying motivations of these two directions as well as our primary reflections on how to proceed them.

7.2.1   Group-Specific Knowledge Discovery

With the current system, while being members of an open group, the users are able to:

- Share with the group interesting information that they have published on their different social networks;

- Share with the group interesting information that they have received from their different social networks;

- Collectively define the group's topics of interest, for example, by suggesting new topics, by enriching current topics, by accepting suitable topics, and by ignoring irrelevant topics;

- Contribute to improving the group's contents of interest, for example, by deleting irrelevant contents, by voting contents, by associating additional topics to a content, and by detecting repeated contents.

These activities of the group's members have an explicit and positive consequence on its information sharing process. More specially, they empower the group's reliable sources of information, and make sure that the group is maintained with the good and enriched topics of interests, and the relevant contents.

Besides this sharing purpose, the members' activities also have an indirect benefit. They actually generate and make available secondary data, which, if correctly analysed, could unveil important and strategic information and knowledge to a given group. In the chapter *Web-based Prototype*, we have seen the *insight* components, which graphically show some interesting information about a group such as the evolution of the topics over time, the repartition of the topics, and the popular topics with their most active contributors. These statistic-based elements are only some simple examples of group-specific knowledge, which can be extracted from the members' activities. Other more sophisticated types of group-specific knowledge can include the knowledge on the members' affinities, their respective expertises on different topics, the group's trending topics, and so forth. All of this group-specific knowledge provides a synthesised and clarified vision on the group, which may be served as a base for following the evolution of the participation of the members and the domain of interest of the group. Below, we dig a little bit deeper into some interesting analysis and their corresponding representation forms.

7.2.1.1 Computational Analysis

Interest Profiling

Within a group, it is important to know who are interested in a given domain (i.e. subject/topic/area) and whether with a high or medium or low degree of interest. Indeed, such knowledge allows to easily and quickly determine the members who most likely have the good answer to a particular question or issue.

With the proposed data model, it is direct and easy to know which members are following a given topic of interest. Also, it is not difficult to compute how many contents of interest a given member contributes to the topic. Although, this information is correlated with the member's degree of interest in the topic in question, it alone is not sufficient to allow to measure the real interest degree. Other interesting factors can be taken into consideration for profiling (i.e. weighting) the member's interest on various topics. For example, the number of contents of interest originated from the member's social data can be split into

the number of those published by the member and the number of those published by the member's social friends. The former number is obviously more significant than the latter. Further, the actions of creating a new topic or of enriching a current topic with additional selectors can be considered as more important than the action of simply accepting a topic. Likely, the actions that the member has taken on the contents of interest (e.g. delete, vote, tag, report, etc.) related to a topic, can be used as proof of the member's expertise on the topic.

The temporal dimension is another important factor to consider when profiling the member's interests [5, 113]. It is logical to give higher weight for interests occurred recently, and lower weight for older interests.

Group Connectedness

The connectedness between two members of a group is a computational measure showing how close and similar the two members are. The group's connectedness is thus the aggregation of all possible weighted connections between its members. This metric is important to the group, as it reflects the structure and the strength of its internal collaboration. It can, for example, be used to select a subset of members who likely work efficiently together within a given project.

The connectedness between two members can be computed taking into consideration their social proximity as well as their similarity [72]. In our case, the social proximity between two members can be determined by whether or not the two members are connected on social networks, and/or how many common social friends they have. The similarity, especially the interest similarity between two members can be derived from the similarity of their two respective interest profiles.

Trending Topics

When a group grows in size, and the number of topics of interest and the number of retrieved contents increase, it becomes difficult to follow the evolution of all topics. Therefore, it will be interesting to know the trending topics, thus paying more attention to them.

The number of recently retrieved contents is an explicit and quite good indicator to determine whether or not a topic is popular. Nevertheless, to better identify the trending topics, this indicator can be completed by additional indicators such as the creation time of the topic, the number of members who have followed the topic, the number of selectors associated to it, and so on, during a certain observation time.

Social Network Cloud Computing **Social Media**
Collective Intelligence Big Data Social Marketing
Internet of Things Social Television **Community Management**

FIGURE 7.1: Trending Topic Cloud

7.2.1.2 Information Visualization

The outputs of the aforementioned analysis are abstract data, which are mostly numerical and not really intuitive for the end users to apprehend. Therefore, they need to be put into another more synthetic representation form in order to reinforce the users' cognition. *Information visualization* [142] is one of the best options to that purpose. According to the supposed output of each computational analysis, there may be several different visual representations. Below are some representative examples.

For the trending topics, we can use the *tag cloud*, in which tags are the topic names, and the importance of each topic is shown with the tag font size or color. For example, with the topic cloud illustrated in Figure 7.1, we immediately understand that "big data", "social media", and "social network" are the three trending and important topics of the group in question.

To visually represent the group connectedness, we can use an undirected graph, in which each node is a member. Two nodes are linked together when the corresponding members' connectedness score is positive (or higher than a certain threshold). The connectedness score furthermore determines the thickness of the link in question. The bigger the connected score, the thicker the link. To help with visualizing the graph, we can furthermore apply a suitable clustering algorithm on the graph in such a way that several sub-groups of members are more visible as illustrated in Figure 7.2.

In addition to these two top-level visualizations, we can also consider some more specified visualizations which allow to focus on a given member or a given topic. The three graphs shown in Figure 7.3 are interesting examples. The first graph includes at its center a given topic, which is surrounded by the nodes representing the members interested by the topic. The closer a member is to the topic, the more interested and specialized he/she is in the topic. Following the same principle, the second and third graphs show the connectedness of a given member with other members and his/her interest degrees with respect to the different topics. In the second graph, the closer a member is to the member in question,
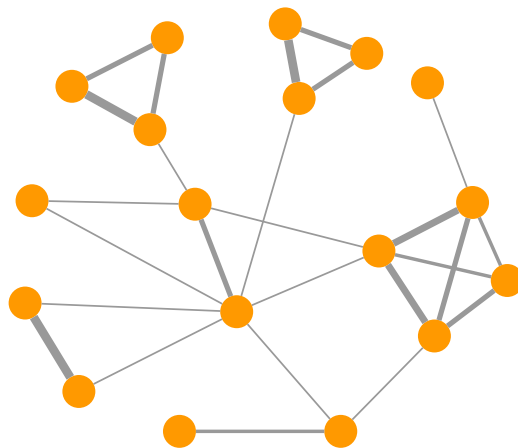
FIGURE 7.2: Member Graph

the more connected they are. In the third graph, the closer a topic is to the member, the more interested and specialized the member is in the topic.

Note that the aforementioned computational analysis and graphical representations are just some examples of the group-specific knowledge discovery based on the data made available by the contributions and exchanges of the group members. Other interesting metrics and graphical format, such as those proposed by Sato and Barthès [130] for following the evolution of the participation and the domain of a community of interest, are also worth considering for this perspective.

7.2.2 Distributed Architecture

Our approach is at the moment proposed with a centralized architecture where the social data of all users are aggregated and stored within a unique server, and where the provided services are accessible via a Web-based application. Such configuration does not require a lot of time and specific technologies to be deployed, and is easy to be tailored and customized. The users do not need specific knowledge and additional tools to utilize the system. Additionally, it makes it possible to further analyse the group members' activities with the objective to discover group-specific knowledge as introduced in the previous subsection.

However, this centralized configuration presents some considerable limitations. Firstly, even though the social data are in principle exclusive to and can be deleted at any time by their owners, some users may still be worried about the privacy and the security of their social data kept within a remote place. This privacy concern is even worsened in the case that an enterprise builds a platform, which also aggregates the social data of its internal collaborators. Especially, some of the collaborators are users of SoCoSys as well. Given
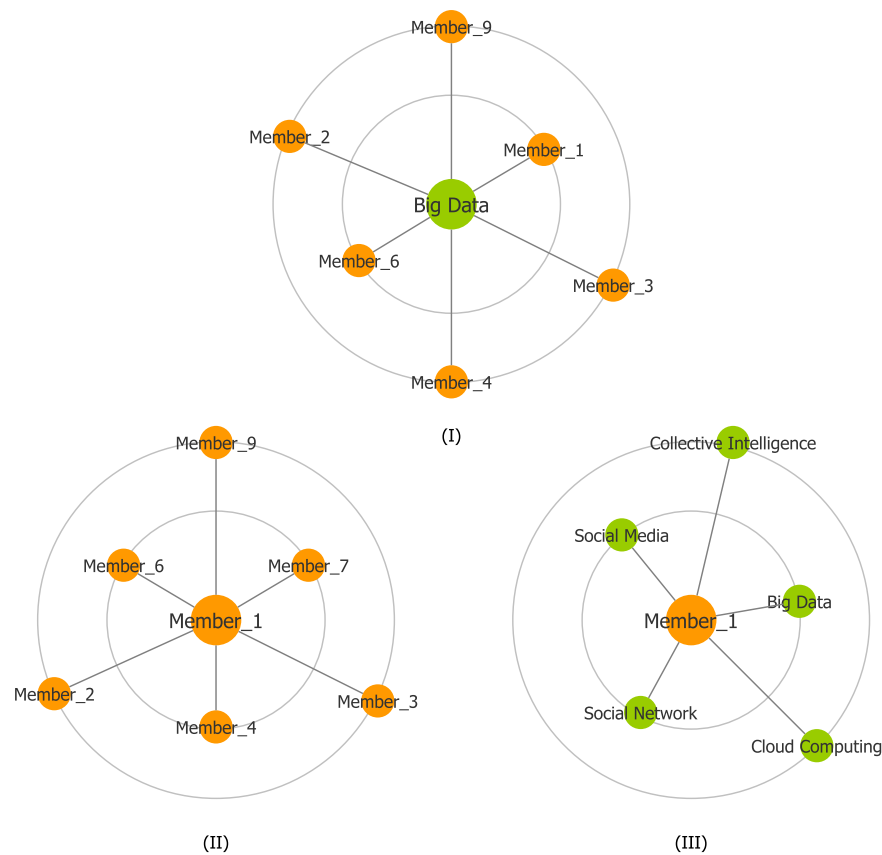
FIGURE 7.3: Focused Visualization

the lack of interoperability between the two systems, the social data of these users will be duplicated and maintained within two different places. Secondly, there may be a volumetric issue, since all the users' social data are inserted into, analysed within, and queried from a single server. This server may be overwhelmed and even crashed upon a growing demand, which probably leads to the temporary or definitive loss of all or part of the retrieved contents of interest.

7.2.2.1   A Distributed Scenario

The aforementioned issues, typical of centralized systems, have long been outlined and addressed in many works. An obvious and direct solution is to decentralize the system into a distributed architecture like a *Peer-to-Peer* network, in which the participants share resources amongst each other without passing the intermediary entities [131]. Such Peer-to-Peer paradigm has successfully been used in many application domains, such as for indexing and searching documents in personal and collective memories [90], for enabling the distributed collaborative content editing [114], and recently for building *Distributed Social Networks* (or *Federated Social Networks*) [34, 115, 133].
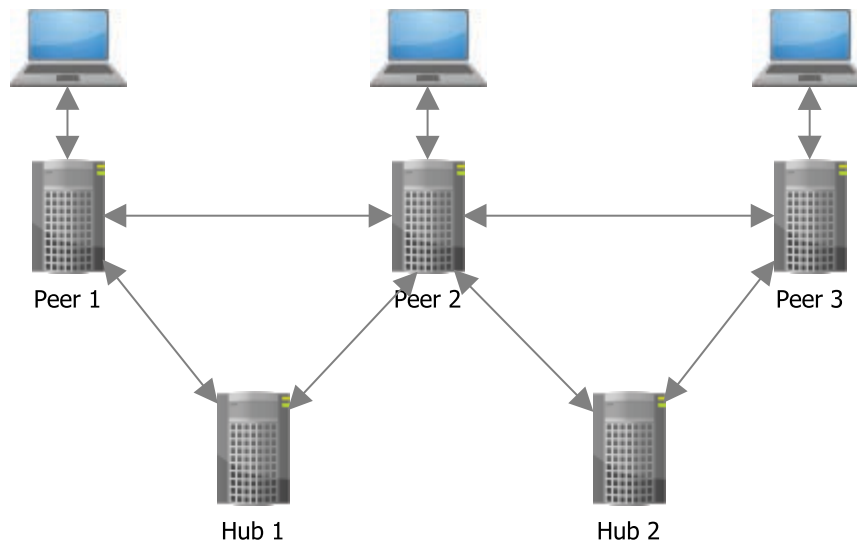
FIGURE 7.4: A possible distributed configuration

Based on the Peer-to-Peer social networks introduced in [9, 115, 147], and taking into consideration the characteristics of our user-centered and group-based approach, we think a scenario, roughly illustrated in Figure 7.4, is interesting and worthy of further exploration.

In this scenario, there are two major actors which are *peers* and *hubs*. A peer represents a trusted server that a given user has chosen to host his/her social data and contents of interest. On this peer, the user thereby needs to install a personal version of SoCoSys, which aggregates the user's social data from different social networks, and extracts the contents of interest.

A hub playing an intermediary role, needs to be deployed in a server. It stores neither the user's social data nor the group's contents of interest, but contains a list of open group references (i.e. name, members, topics/selectors) that the users can look for, subscribe to or unsubscribe from. A user can create a group within his/her local SoCoSys and push it to a given hub so that it is appended to the list of open groups. Otherwise, the group remains private and closed within the user's local SoCoSys. When subscribing to an open group, the user by default follows all of its topics of interest and their associated selectors. Of course, the user can later personalize his/her topics of interest from his/her local SoCoSys. In addition, the user can locally add new topics and/or additional selectors, which will be pushed to the hub to update the topics/selectors of the group concerned. The hub will then spread the update to all the registered members of the group so that they can also update their local topics/selectors.

When a peer p_1 discovers new contents of interest from the social data of its owner, the user u_1, it will notify the corresponding hub about the discovery without unveiling the

contents. Given the target group, the hub will notify the registered member peers about the incoming contents (i.e. their source p_1 and their original identifiers). If the contents are of the interest of the user u_2, the corresponding peer p_2 will request the peer p_1 for the full contents by providing their identifiers. If it is the first time that the peer p_2 asks the peer p_1 for contents, it must first provide some authentication proofs, based on which the user u_1 can decide whether or not to trust the user u_2. Once trusted, the peer p_2 will receive from the peer p_1 the full contents. The contents will then be saved in the local storage unit of the peer p_2, thus being available to the user u_2.

With this scenario, the users have a total control over their aggregated social data as well as the extracted contents and the people with whom they share. The scalability is furthermore no longer a critical performance factor, since each local SoCoSys has to aggregate and analyse only a small quantity of social data.

7.2.2.2   A Semantic Distributed Scenario

The previous scenario is possible if, and only if, every peer uses a same version of SoCoSys, and subsequently a common data representation model. If a peer decides to modify its data representation model, it will cause a compatibility problem for other peers who receive its contents of interest.

Therefore, it is interesting to extend the previous distributed scenario to a semantic distributed scenario based on the *Linked Data* principles [146]. More specially, the current relational data model will be mapped to the RDF model [140]. The contents of interest will furthermore be identified by dereferenceable URIs and be made available via the different endpoints exposed by the different peers.

This semantic distributed scenario not only ensures a better data interoperability between the different peers, but also has other important advantages. Firstly, it is not necessary to replicate a content of interest within the different peers. A peer only needs to refer to the corresponding content stored in another peer using its URI. The full content can easily be retrieved by SPARQL queries at any time. Secondly, it is possible to link the RDF data to the *Web of Data* [22] so that we can discover more contents of interest.

Chapter 8

Summary

Contents

The social media have played an increasingly important role in many areas of our every day life. They include a wide range of services that exist through various constantly evolving forms. Among others, social network sites such as Facebook, LinkedIn, Twitter and Google+ have recently exploded in popularity. These open and large-scale social networking services attract millions of users from around the world, who communicate with each other, share and publish information and contents at an unprecedented rate. Alongside their benefits, social network sites have also raised various issues and challenges, some of which are very complicated and require multi-discipline approaches and solutions. In this thesis, we addressed two particular problems, which are *information overload* and *"walled gardens"*. These two problems, typical of today social network sites, prevent the users from fully exploiting and benefiting from the wealth of information available on social network sites. The users have a lot of difficulties to filter all incoming information, to discover additional information from outside of their friend cycles, and importantly to share the interesting contents with their different groups of interest. We therefore proposed a novel and unique approach helping the users to overcome such difficulties. This chapter summarizes the objectives, the main contributions, and the perspectives of this work.

8.1   Summary of Objectives

The two problems of information overload and "walled gardens" led to a number of subsequent consequences. Firstly, the user is often overwhelmed by the huge number of incoming information, which is scattered across different social networks sites. The user can not spend a lot of time and efforts to manually extract contents of interest from all incoming information. Many contents of interest are therefore missed by the user.

Secondly, most social network sites establish privacy policies restricting what a user may receive within his/her social streams. The user typically receives the contents shared by his/her social connections, who must also be the members of the same social network site. The user can add new friends to expand his/her information sources, but at the risk of increasing the chance of information overload.

Thirdly, the user can be a member of different groups of interest, and the information that he/she published on a particular social network site, may interest other members of one of his/her groups. However, there is no guarantee that all the members of a group are connected to a same social network and connected to each other to be able to receive interesting information shared by one of them. There had not been an efficient solution for a group of interest to tap into the contents published by its members across different social network sites to retrieve some parts relevant to its topics of interest, except to ask each member to make extra efforts to copy the contents of interest into the group.

Taken into consideration these consequences, we asked the three questions:

1. *The filtering question*: How to help users to extract contents of interest from their different social networks with less effort and without altering their social networking experience ?

2. *The discovering question*: How to help the users to discover additional contents of interest from outside of their cycles of friends ?

3. *The sharing question*: How to make it possible for the users to share the contents of their various social streams with their respective groups without extra charge ?

Although there are many works attempting to address the two initial problems, none of them includes enough features to answer all the three asked questions. Especially, we did not find out any work studying the sharing question. Given that lack of a complete and unified solution able to answer all the three questions, our work was therefore aimed at researching for such a solution. Furthermore, we intended to achieve the proposed solution by building a working prototype (i.e. proof of concept), which would be served as a "testing

ground" for our assumptions on the social network sites and the expected benefits of our proposed solution.

8.2 Summary of Contributions

In answering the aforementioned questions, we have proposed a novel *User-centered and group-based approach for social data filtering and sharing*. In this section, we recall its different aspects including its conceptual design, its baseline modular system architecture, and its Web-based prototype, which are the main contributions of this work.

8.2.1 A Conceptual Design

Our proposed approach consists of two main components: (1) *User-centered social data filtering*, and (2) *Group-based social data sharing*. The first component answers the filtering question by allowing the user to aggregate their different social streams and to extract contents which are of the user's interests. The second component answers both discovering and sharing questions by enabling collaborative spaces where the members of a given group of interest can share with each other the contents of interest extracted from their respective aggregated social data, thus accessing to more interesting contents. In general, there are three important conceptual elements to note.

Firstly, we have built an adapted common model based on FOAF and ActivityStream. This generic model is able to integrate the most frequent information dimensions of social data available in popular social network sites. Especially, it is easy to be extended to include new types of social data.

Secondly, for filtering and organizing contents of interest, we have applied a group - topic - selector structure. A group, whether private or open, contains a number of topics of interest, each of which is technically specified by one or several selectors. With such organization, the contents of interest, matched and extracted by the selectors from the user's aggregated social data, are split into different groups, and are assigned to various topics. This way, the user can easily access to the expected contents by selecting the corresponding group and topic.

Thirdly and most importantly, we have based the two components on a user-centric design. More specially, the user is asked to authenticate and to authorize access to his/her social data across social network sites, thus being free to choose which social account to be aggregated. The user is also asked to explicitly and gradually add his/her topics of interest and associate to them the appropriate filtering techniques (i.e. selectors). This way, the

system knows exactly what the user want, and provides the user with better results. As a member of a given group, the user is moreover able to decide which part of his/her aggregated social data should be open, processed and eventually shared with the group, to prevent undesired information from being unveiled. Especially, the members of a group are encouraged to contribute to defining the group's topics of interest so that the group is maintained with the good and enriched topics of interests, and the relevant contents.

8.2.2   A Baseline Modular System Architecture

To achieve the conceptual design, we have presented a baseline technical solution, which has a centralized modular architecture. This architecture is composed of three main components: *aggregating component*, *searching component* and *collaborative component*. Each of these three components contains different modules, and has the specified roles and functions.

The aggregating component is responsible for aggregating and storing the users' social data from different social network sites. It is straightforwardly based on a variety of APIs provided by the social network providers to retrieve the users' social data and uses the hand-crafted rules to map the social data with the common model. It moreover enriches the aggregated social data by extending them with the contents from the external resources.

The searching component is responsible for extracting contents of interest. We have mainly implemented this component using the open source Lucene platform, which is considered for providing a robust and scalable indexing and retrieval platform. More specially, Lucene is used to index the enriched aggregated social data and to search the resulting indexes against the user's specified selectors, which can be keywords or hashtags or concepts.

The collaborative component has been added to allow the user to perform and benefit from other more explicit collaborative efforts such as deleting irrelevant contents, promoting relevant content, or demoting unsuitable contents, adding additional topics to a given content and so forth. Such efforts can fill the potential gaps of the current searching performance.

For each containing module of the three components, we have also discussed its possible issues and improvements.

8.2.3   A Tested Web-based Prototype

Based on the proposed system, we have built our first prototype, called SoCoSys supporting the three very popular social network sites, namely Facebook, Twitter, and LinkedIn. This

prototype has actually been implemented as a Web-based application so that the end users are able to access to it from anywhere without restriction. Despite their simplicity due to the objective of facilitating the user's tasks, the user interfaces of SoCoSys have fully complied with the conceptual requirements. Some new group-specific, namely viewing members and getting insights, have even been added to provide the group with primary means for accessing to the advanced knowledge on its internal collaboration.

Using this prototype, we have carried out two small tests with two different test groups. The first test group consisted of international PhD students at the University of Technology of Compiègne, whereas the second test group consisted of engineering apprentices at the same university. The analysis based on the data obtained from the first test, and the responses to the questionnaires of the participants of the second test, have confirmed that the two addressed problems are real, and that social network sites represent potential sources of information and contents of multiple domains. In addition, the participants' positive opinions on the use of SoCoSys have approved its utilities, its usability, and its functionalities.

8.2.4 Comparative Discussions

Social network aggregators :   SoCoSys is a social network aggregator, as it helps the users aggregate their social data from the three popular social networks (i.e. Facebook, Twitter, and LinkedIn) and provides them with a single access to those data. However, unlike the current commercial social network aggregators which try to pull nearly all kinds of things happened on the user's different social streams together so that he/she can read, share and comment them without leaving the platform, SoCoSys attempts to organize (i.e. filter, index) the most informative part of those things. SoCoSys therefore relies on a common model (based on FOAF and ActivityStream) to retrieve from the social networks only the needed social data. Using SoCoSys does not means that the user has to stop using his/her current social networks. The user keeps using his/her social networks normally, but with less effort, in less time, and with greater efficiency when extracting helpful contents.

Filtering solutions :   Our approach took a semi-automated way for filtering the social data. The system extracts from the social data the contents relevant to the topics of interest explicitly defined by the users. This method requires less effort than other completely manual methods (e.g. *Manual friend grouping*). It gives the user some control on the filtering process. First, the user does not have to follow some limited and predefined topics (i.e. *Content classification*), or to follow too many topics generated by the system (i.e. *Topic detection*), but can personalize his/her topics of interest over time. Second, unlike the less transparent methods like *Personalized filtering* or *Stream (re)ranking*, the user

has a clear idea why a piece of content is selected as relevant. The filtering component is build based on the *Information Retrieval* generic techniques, is then very extensible and suitable for all the social data, mostly textual data, from different social network sites. To reduce the users' extra effort when defining the topics and also the selectors, our approach furthermore proposed the users to collaborate within groups so that a relevant topic or a good selector suggested by a member can inspire others.

Collaborative systems :   SoCoSys can also be considered as a collaborative system, especially for information sharing. It allows the users to share with their respective groups of interest the helpful contents originated from their different social streams in an effortless manner. While other systems require their users to manually select and copy the contents inside their groups, SoCoSys automatically extracts the contents from the group members' social data, obviously with their permissions. In our approach, the group's collective source number is proportional to the size of its members and their circles of friends. Such collective source is more dynamic and divers than an official source (e.g. news articles), and is more targeted than an entire social network (e.g. all public messages published on the social network). Finally, the members of a group are able to carry out an objective-driven collective action (e.g. technological watch) by using the Hashtag method to retrieve the related information and news.

8.3   Summary of Perspectives

The developed system is, at this stage of the project, only a proof of concept demonstrating the expected benefits (i.e. filtering and sharing) of our proposed approach. It has been furthermore implemented with a centralized architecture, some modules of which are for now employing generic and simplified techniques, thus being improvable. Besides, the experimentation with two small test groups has not allowed us to properly measure the effectiveness and the scalability of the system. Taking all of this into consideration, we set out different interesting perspectives for future work.

In short term, we will incrementally improve and test the system with other bigger groups of users, possibly from a given organization (i.e. enterprise). These future tests will not only provide us with new sets of data large enough for a complete evaluation on the system effectiveness and scalability, but also allow us to explore the new uses of our system.

For the long-term perspectives, we envision two independent directions, which will review some fundamental aspects of our proposed approach. The first direction *Group-specific knowledge discover* will attempt to extend the initial scope of the approach (i.e. filtering

and sharing). It will look for different computational analysis on the members' activities, and for suitable representation forms to make it possible to discover some group-specific knowledge, for example the group connectedness, the trending topics, the member topic expertise.

The second direction *Distributed architecture* will examine the possibility of transforming the system architecture from the current centralized configuration to a new distributed configuration. In such a distributed configuration, the user will have a total control over his/her aggregated social data as well as the extracted contents and the people with whom he/she shares. The scalability will be furthermore no longer a critical performance factor, since the whole process will be not performed on a single machine, but distributed over many machines.

Appendix A

The Conversation Prism

The Conversation Prism[1] is a visual map of the social media landscape. It is an ongoing study in digital ethnography that tracks dominant and promising social networks and organizes them by how they are used in everyday life. Every year, Brain Solis and JESS4 review and re-adapt the prism by removing disappeared services and adding new ones.

Social media services are evolving rapidly and the landscape continues to grow. When Brian Solis introduced the first Conversation Prism in 2008, the world was a seemingly simpler place. There were 22 social media categories, each of which had just a handful of brands. In 2013, the latest Conversation Prism was released and has four additional categories with at least six brands in each (see Figure A.1).

---

[1] https://conversationprism.com/
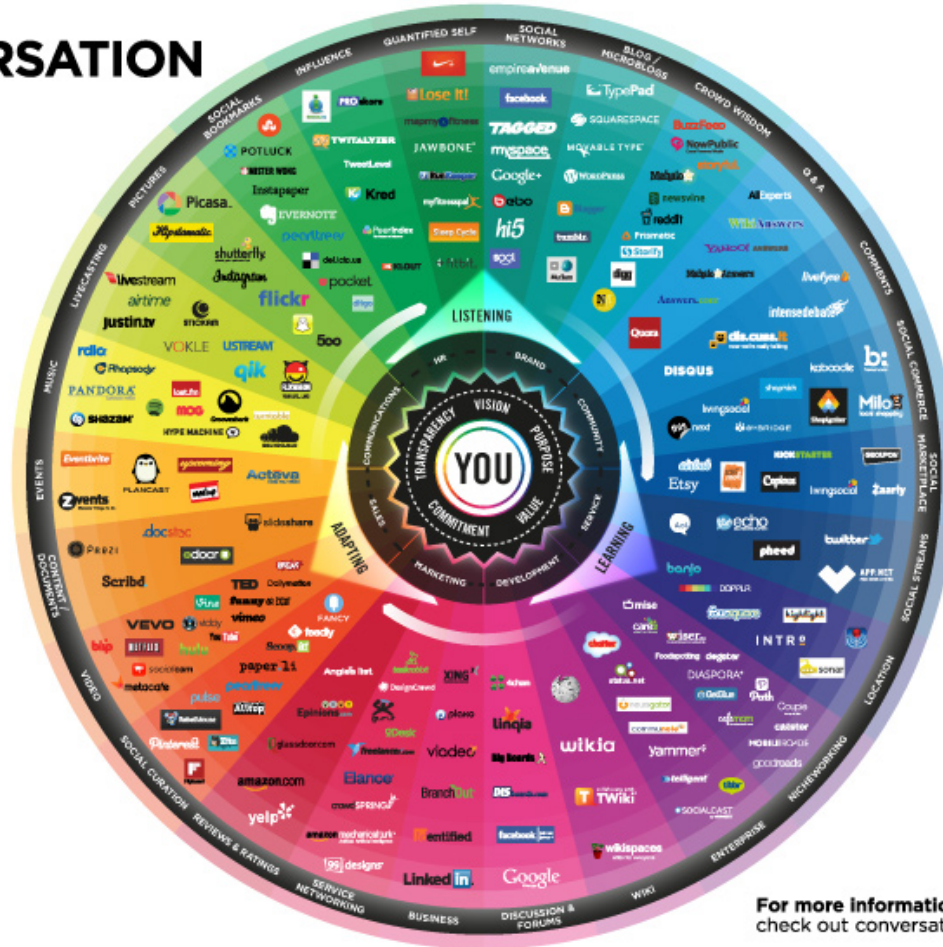
FIGURE A.1: The Conversation Prism by Brian Solis and JESS3

Appendix B

MySQL Physical Schema

As mentioned in Section *Aggregating Part* of Chapter *Technical Solution*, we have at the moment chosen to use a relational database, namely MySQL, for the storing purpose. Here, we can see two parts of the database schema. The first part illustrated by Figure B.1 is necessary for storing the users' aggregated social data. The second part illustrated by Figure B.2 is devoted to group settings.

Both parts comply with the conceptual requirements defined in Chapter *Conceptual Design*. Most of classes and class members and relationships are maintained and represented by the corresponding tables often with the same names (see Table B.1). Besides, constraints including primary keys, foreign keys, other unique keys, and check constraints are added.

For performance reasons, there are nevertheless a couple of adjustments that are quite important to note. Firstly, we have directly associated the two classes *SocialActivity* and *SocialData* together within the table *social_data* (see Figure B.1). Secondly, we have created a unique table *selector* for the class *Selector* and its entire hierarchy (see Figure B.2). Also, we have added the table *query* to save the expanded queries as early as they are built in order to reuse them later. The tables *tag* and *vote* have been added to store the members' votes (up or down) and additional topics for the contents shared within a given group (see the subsection Enhancement of Chapter 4).

TABLE B.1: Data dictionary

Table	Conceptual element	Description
user	User	SoCoSys user accounts
social_profile	SocialAccount	the social accounts that the users linked to their SoCoSys accounts
social_network	SocialNetwork	the original social network of the social accounts and the social data
social_data	SocialActivity ⊕ SocialData	the users' social data aggregated from the different social networks
social_interest	SocialActivity("add") ⊕ SocialData(Interest)	the users' interests
social_friend	SocialActivity("befriend") ⊕ SocialData(Member)	the users' social contacts
social_post	SocialActivity("post") ⊕ SocialData(Post)	the contents published by the users
social_following_post	SocialActivity("receive") ⊕ SocialData(Post ⊕ Member)	the contents shared with the users
group	Group(OpenGroup ⊕ PrivateGroup)	the collaborative spaces where the member(s) can organize and/or share their own social data
user_group	memberOf	the users' memberships and sharing settings with respect to their different groups
topic	Topic	the group's topics of interest
user_topic	User *follows* Topic	the users' personalized topics of interest within a given group
selector	Selector(Keyword ⊕ Hashtag ⊕ Concept)	the selectors of a given topic
user_selector	User *accepts* Selector	the users' personalized selectors of a following topic
query		the final and extended query of a given selector
content	Content	the relevant contents extracted from the members' shared social data
content_selector	Content *matches* Selector	a content retrieved when matching at least one of the selectors of a given topic
vote		the members' votes (up or down) for the contents shared within a given group
tag		the additional topics manually assigned to a given content by the members within a given group
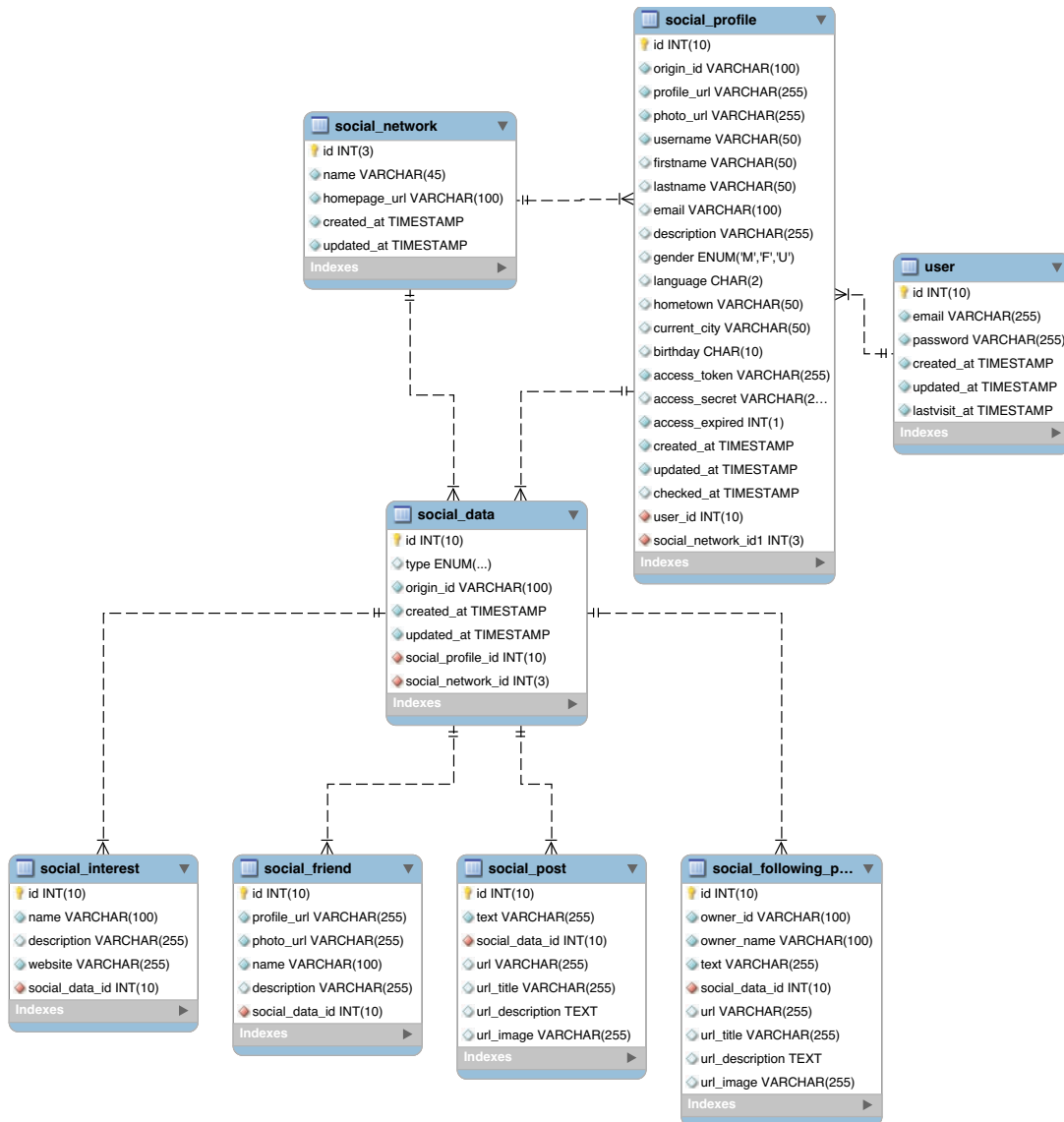
FIGURE B.1: The tables necessary for storing social data
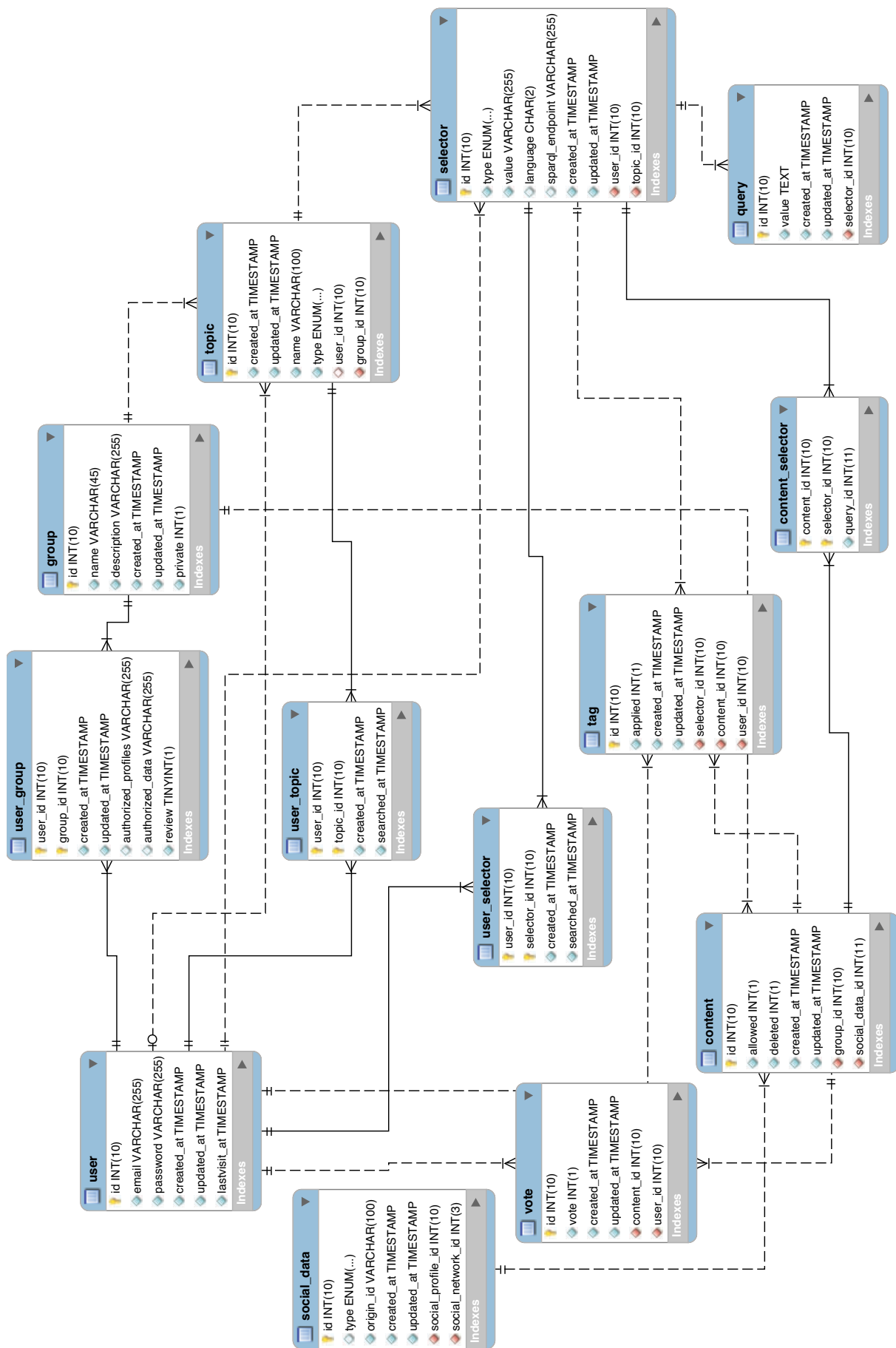
FIGURE B.2: The tables devoted to group settings

Appendix C

French Questionnaires

Questionnaire de l'usage de SoCoSys

Ce questionnaire s'inscrit dans le cadre d'un projet doctoral au sein du laboratoire Heudiasyc (UTC). Il est composé d'une vingtaine de questions. Vos réponses, totalement anonymisées, vont nous aider à mieux évaluer notre travail.
Merci de prendre quelques minutes pour remplir ce questionnaire,

I. Information personnelle

1. **Sexe :**
   Mark only one oval.

   ◯ Femme

   ◯ Homme

2. **Âge :**

   ....

II. Usage des réseaux sociaux

3. **Avez vous au moins un profil sur les réseaux suivants ?**
   Tick all that apply.

   ☐ Facebook

   ☐ Twitter

   ☐ LinkedIn

4. **Si vous utilisez plusieurs réseaux sociaux, quel est votre habitude ?**
   Mark only one oval.

   ◯ Vous essayez de les fréquenter tous

   ◯ Vous fréquentez un réseau principal et visitez de temps en temps les autres

5. **Quelle est votre fréquentation actuelle d'utilisation des réseaux sociaux ?**

Mark only one oval.

◯ Plusieurs fois pendant la journée

◯ Au moins une fois par jour

◯ Quelques fois par semaine

◯ De temps en temps

6. **Partagez vous des informations d'intérêt sur les réseaux sociaux ?**

Informations d'intérêt sont ceux qui vous intéressent et/ou intéressent vos amis.

Tick all that apply.

☐ Pas du tout

☐ Un peu

☐ Beaucoup

7. **Pensez vous que les réseaux sociaux vous apportent beaucoup d'informations intéressants ?**

Tick all that apply.

☐ Pas du tout

☐ Un peu

☐ Beaucoup

8. **Pensez vous qu'il y a trop d'informations, intéressantes et pas intéressantes, qui vous parviennent sur les réseaux sociaux ?**

Mark only one oval.

◯ Oui

◯ Non

III. SoCoSys

9. **Trouvez vous utile d'agréger vos réseaux sociaux, d'en extraire des bonnes informations, et de les rendre accessibles à un endroit unique ?**

Mark only one oval.

◯ Oui

◯ Non

◯ Peut être

10. **Trouvez vous intéressant de partager les informations extraites des vos réseaux sociaux avec vos groupes d'intérêt ?**

Groupes d'intérêt regroupent des personnes partagent de mêmes intérêts
Mark only one oval.

◯ Oui

◯ Non

◯ Peut être

11. **Pensez vous que SoCoSys offre bien ces deux fonctionnalités ?**

Mark only one oval.

◯ Oui

◯ Non

◯ Peut être

12. **Est il en général facile d'utiliser les interfaces Web de SoCoSys ?**

Mark only one oval.

◯ Oui

◯ Non

13. **Est il en général simple de comprendre le fonctionnement de SoCoSys ?**

Mark only one oval.

◯ Oui

◯ Non

14. **Etes vous d'accord avec l'organisation par groupes (privé vs. ouvert) ?**

Mark only one oval.

◯ Oui

◯ Non

15. **D'après vous, le mécanisme de filtrage actuel est-il pertinent ?**

Vous entrez manuellement vos propres sujets d'intérêt, le reste du processus est automatisé.
Mark only one oval.

◯ Oui

◯ Non (il faut tout automatiser)

16. **Etes vous d'accord avec la possibilité de limiter ce que l'on peut partager dans un groupe ouvert ?**
Vous pouvez choisir quelles parties de vos informations à éventuellement partager avec le groupe en utilisant les paramètres de partage.
Mark only one oval.

( ) Oui

( ) Non (il faut que tous les membres ouvrent tous)

17. **Etes vous d'accord avec la définition collective des sujets d'intérêt dans un groupe ouvert ?**
Tous les membres peuvent proposer les sujets d'intérêt.
Mark only one oval.

( ) Oui

( ) Non (il faut qu'un seul modérateur puisse le faire)

18. **Etes vous d'accord avec la personnalisation des sujets d'intérêt dans un groupe ouvert ?**
Vous pouvez accepter ou ignorer certains sujets.
Mark only one oval.

( ) Oui

( ) Non

19. **D'après vous, trois méthodes de filtrage sont-ils suffisants ?**
Hashtags, Keywords, Concepts
Mark only one oval.

( ) Oui

( ) Non

20. **A votre avis, SoCoSys doit être plus tôt destiné à :**
Mark only one oval.

( ) Un usage personnel

( ) Un usage collectif

( ) Les deux

21. **Pensez vous que SoCoSys peut être également mis en place au sein des organisations/entreprises comme un outil collaboratif de travail ?**
Mark only one oval.

( ) Oui

( ) Non

22. **Finallement, souhaitez vous poursuivre l'utilisation de SoCoSys après le test ?**

Mark only one oval.

- ◯ Oui
- ◯ Non
- ◯ Peut être

23. **Si non, pour quelle raison ?**

....

....

....

....

....

Suggestion

24. **Pensez vous que SoCoSys doit être encore amélioré ?**

Mark only one oval.

- ◯ Oui
- ◯ Non

25. **Si oui, avez vous des suggestions à propos des interfaces de SoCoSys ?**

....

....

....

....

....

26. **Et avez vous des suggestions à propos du fonctionnement de SoCoSys ?**

....

....

....

....

....

27. **Toutes autres suggestions/remarques ?**

      ....

      ....

      ....

      ....

      ....

Merci d'avoir rempli ce questionnaire

Appendix D

Publications

D.1   Journal Articles

- Vu, X. T., Abel, M.-H., and Morizet-Mahoudeaux, P. An aggregation model of online social networks to support group decision-making. *Journal of Decision Systems*, 23, 1 (2014), pp. 24-39.

- Vu, X. T., Abel, M.-H., and Morizet-mahoudeaux, P. A User-centered and Group-based Approach for Social Data Filtering and Sharing. *Computers in Human Behavior Journal*, 2014, doi:10.1016/j.chb.2014.11.079

- Vu, X. T., Abel, M.-H., and Morizet-Mahoudeaux, P. A User-centered Model for Integrating User Social Data into Communities of Interest. Journal of Data & Knowledge Engineering, 2015, doi:http://dx.doi.org/10.1016/j.datak.2015.04.004

D.2   Conference Proceedings

- Vu, X. T., Abel, M.-H., and Morizet-Mahoudeaux, P. An Aggregation Model of Online Social Networks to Contribute to Organizational Knowledge Management. In *Proceedings of the 1st International Conference on Knowledge Management, Information and Knowledge Systems* (Hammamet, Tunisia, 2013), ISTE-Wiley, pp. 25-37.

- Vu, X. T., Abel, M.-H., and Morizet-mahoudeaux, P. Empowering Collaborative Intelligence by the use of User-centered Social Network Aggregation. In *Proceedings of the 2013 IEEE/WIC/ACM International Conference on Web Intelligence WI 2013* (Atlanta, Georgia, USA, 2013), IEEE Computer Society, pp. 425-430

D.3 Book Chapters

- Vu, X., Abel, M.-H., and Morizet-Mahoudeaux, P. Integrating social network data for empowering collaborative systems. In *Knowledge and Systems Engineering*, vol. 245 of *Advances in Intelligent Systems and Computing*. Springer International Publishing (2014), pp. 109-119.

- Vu, X. T., Abel, M.-H., and Morizet-Mahoudeaux, P. Social networks: Leveraging user social data to empower collective intelligence. In *Information Systems for Knowledge Management* (2014), John Wiley & Sons, Inc., pp. 33-60.

D.4 Poster

- Vu, X. T., Morizet-mahoudeaux, P., and Abel, M.-H.. User-centered social network profiles integration. In *Proceedings of the 9th International Conference on Web Information Systems and Technologies* (Aachen, Germany, 2013), SciTePress, pp. 473-477.

Bibliography

[1] AARON, S. 6 new facts about facebook, Feb. 2014. Available at http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/.

[2] ABDEL-HAFEZ, A., AND XU, Y. A Survey of User Modelling in Social Media Websites. *Computer and Information Science 6*, 4 (Sept. 2013), 59–71.

[3] ABDESSLEM, F., PARRIS, I., AND HENDERSON, T. Reliable online social network data collection. In *Computational Social Networks*, A. Abraham, Ed. Springer London, 2012, pp. 183–210.

[4] ABEL, F., CELIK, I., HOUBEN, G. J., AND SIEHNDEL, P. Leveraging the semantics of tweets for adaptive faceted search on twitter. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2011), vol. 7031 LNCS, pp. 1–17.

[5] ABEL, F., GAO, Q., HOUBEN, G., AND TAO, K. Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *Proceedings of the ACM WebSci'11* (Koblenz, Germany, 2011), pp. 1–8.

[6] ABEL, F., GAO, Q., HOUBEN, G.-J. G., AND TAO, K. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th international conference on User modeling, adaption, and personalization* (Berlin, Heidelberg, 2011), UMAP'11, Springer-Verlag, pp. 1–12.

[7] ABEL, F., HANNOVER, A. D., HENZE, N., KRAUSE, D., HERDER, E., AND KRAUSE, D. Linkage, Aggregation, Alignment and Enrichment of Public User Profiles with Mypes. In *Proceedings of the 6th International Conference on Semantic Systems* (New York, NY, USA, 2010), I-SEMANTICS '10, ACM, pp. 11:1–11:8.

[8] ABEL, F., HERDER, E., HOUBEN, G.-J., HENZE, N., AND KRAUSE, D. Cross-system user modeling and personalization on the Social Web. *User Modeling and User-Adapted Interaction 23*, 2-3 (Nov. 2012), 8–9.

[9] ACKERMANN, M., HYMON, K., LUDWIG, B., AND WILHELM, K. HelloWorld: An Open Source, Distributed and Secure Social Network. In *Proceedings of the W3C Workshop on the Future of Social Networking* (Barcelona, 2009).

[10] ADOBE. Adobe 2013 mobile consumer survey, Mar. 2013. Available at http://success.adobe.com/assets/en/downloads/whitepaper/35508_mobile_consumer_survey_results_UE_final-2.pdf.

[11] AGARWAL, D., CHEN, B.-C. B., GUPTA, R., HARTMAN, J., HE, Q., IYER, A., KOLAR, S., MA, Y., SHIVASWAMY, P., SINGH, A., AND ZHANG, L. Activity Ranking in LinkedIn Feed. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2014), KDD '14, ACM, pp. 1603–1612.

[12] AGGARWAL, C. An Introduction to Social Network Data Analytics. In *Social Network Data Analytics*, C. C. Aggarwal, Ed. Springer US, 2011, pp. 1–15.

[13] ALEXA. The top 500 sites on the web, 2014. Available at http://www.alexa.com/topsites.

[14] APPELQUIST, D., BRICKLEY, D., CARVAHLO, M., IANNELLA, R., PASSANT, A., PEREY, C., AND STORY, H. A standards-based, open and privacy-aware social web. *W3C Incubator Group Report 6* (2010).

[15] ARMENTANO, M., GODOY, D., AND AMANDI, A. Followee recommendation based on text analysis of micro-blogging activity. *Information Systems* (June 2013).

[16] AUVINEN, A.-M. *Social media - The new power of policial influence.* Toivo Think Tank, 2012.

[17] BECKER, H., NAAMAN, M., AND GRAVANO, L. Beyond trending topics: Real-world event identification on twitter. *ICWSM 11* (2011), 438–441.

[18] BELKIN, N. J., AND CROFT, W. B. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Commun. ACM 35*, 12 (Dec. 1992), 29–38.

[19] BENNETT, S. 92% of companies use social media for recruitment, Oct. 2013. Available at http://www.mediabistro.com/alltwitter/social-media-recruiting_b50575.

[20] BERGER, K., KLIER, J., KLIER, M., AND PROBST, F. A Review of Information Systems Research on Online Social Networks. *Communications of the Association for Information Systems 35*, 1 (2014), 145–172.

[21] BERNSTEIN, M. S. M., SUH, B., HONG, L., CHEN, J., KAIRAM, S., AND CHI, E. H. Eddi: Interactive Topic-based Browsing of Social Status Streams. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2010), UIST '10, ACM, pp. 303–312.

[22] BIZER, C., LEHMANN, J., KOBILAROV, G., AUER, S., BECKER, C., CYGANIAK, R., AND HELLMANN, S. Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web 7*, 3 (2009), 154–165.

[23] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

[24] BOJARS, U., PASSANT, A., AND BRESLIN, J. Data Portability with SIOC and FOAF. In *XTech 2008 conference* (2008), pp. 1–9.

[25] BONCHI, F., CASTILLO, C., GIONIS, A., AND JAIMES, A. Social Network Analysis and Mining for Business Applications. *ACM Trans. Intell. Syst. Technol. 2*, 3 (May 2011), 22:1–22:37.

[26] BONDS-RAACKE, J., AND RAACKE, J. MySpace and Facebook: Identifying dimensions of uses and gratifications for friend networking sites. *Individual Differences Research 8*, 1 (2010), 27–33.

[27] BORGS, C., CHAYES, J., KARRER, B., AND MEEDER, B. Game-theoretic models of information overload in social networks. In *Algorithms and Models for the Web-Graph*, R. Kumar and D. Sivakumar, Eds. Springer Berlin Heidelberg, 2010, pp. 146–161.

[28] BOURKE, S., O'MAHONY, M., RAFTER, R., AND SMYTH, B. Ranking in information streams. In *Proceedings of the companion publication of the 2013 international conference on Intelligent user interfaces companion - IUI '13 Companion* (New York, New York, USA, 2013), ACM Press, pp. 99–100.

[29] BOYD, D. M., AND ELLISON, N. B. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication 13*, 1 (Oct. 2007), 210–230.

[30] BRANDTZG, P., AND HEIM, J. Why people use social networking sites. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5621 LNCS* (2009), 143–152. cited By (since 1996)21.

[31] BRENNER, J., AND SMITH, A. 72% of online adults are social networking site users, Aug. 2013. Available at http://www.pewinternet.org/2013/08/05/72-of-online-adults-are-social-networking-site-users/.

[32] BRICKLEY, D., AND MILLER, L. Foaf vocabulary specification 0.98. *Namespace Document 9* (2012).

[33] BUCCAFURRI, F., LAX, G., NOCERA, A., AND URSINO, D. Moving from social networks to social internetworking scenarios: The crawling perspective. *Information Sciences* (Sept. 2014).

[34] BUCHEGGER, S., SCHIÖBERG, D., VU, L.-H., AND DATTA, A. Peerson: P2p social networking: Early experiences and insights. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems* (New York, NY, USA, 2009), SNS '09, ACM, pp. 46–52.

[35] CAIN-MILLER, C. Another try by google to take on facebook, June 2011. Available at http://www.nytimes.com/2011/06/29/technology/29google.html.

[36] CARMAGNOLA, F., CENA, F., AND GENA, C. User model interoperability: a survey. *User Modeling and User-Adapted Interaction 21*, 3 (Feb. 2011), 285–331.

[37] CARMAGNOLA, F., OSBORNE, F., AND TORRE, I. User data distributed on the social web: how to identify users on different social systems and collecting data about them. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems* (New York, NY, USA, 2010), HetRec '10, ACM, pp. 9–15.

[38] CHEN, J., NAIRN, R., NELSON, L., BERNSTEIN, M., AND CHI, E. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 1185–1194.

[39] CHIU, P.-Y., CHEUNG, C. M., AND LEE, M. K. Online social networks: Why do "we" use facebook? *Computers in Human Behavior 27*, 4 (July 2011), 1337–1343.

[40] CHUN, Y., HWANG, H., AND KIM, C. Development of a Disaster Information Extraction System based on Social Network Services. *International Journal of Multimedia and Ubiquitous Engineering 9*, 1 (2014), 255–264.

[41] COHEN, R., SARDANA, N., RAHIM, K., LAM, D. Y., LI, M., MACCARTHY, O., WOO, E., AND GUO, G. Reducing information overload in social networks through streamlined presentation : a study of content-centric and person-centric contexts Towards a Generalized Algorithm. In *3rd Workshop on Incentives and Trust in E-Communities* (Québec City, Québec, Canada, 2014), S. Marsh, J. Zhang, C. Jensen, Z. Noorian, and Y. Liu, Eds., pp. 13—-18.

[42] DAS, B., AND SAHOO, J. S. Social networking sites–a critical analysis of its impact on personal and social life. *International Journal of Business and Social Science 2*, 14 (2011), 222–228.

[43] DELLA-DORA, L. The state of google+, Feb. 2014. Available at http://wearesocial.net/blog/2014/02/state-google/.

[44] DIMICCO, J., MILLEN, D. R., GEYER, W., DUGAN, C., BROWNHOLTZ, B., MULLER, M., AND STREET, R. Motivations for Social Networking at Work. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA, Nov. 2008), ACM, pp. 711–720.

[45] DUGGAN, M., AND SMITH, A. Social media update 2013, Dec. 2013. Available at http://www.pewinternet.org/2013/12/30/social-media-update-2013/.

[46] ED, O. Google+ sees explosive growth, 50 million-plus users in less than 3 months, Sept. 2011. Available at http://betanews.com/2011/09/26/google-sees-explosive-growth-50-million-plus-users-in-less-than-3-months/.

[47] EPPLER, M. J., AND MENGIS, J. The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The information society 20*, 5 (2004), 325–344.

[48] ESLAMI, M., ALEYASEN, A., ZILOUCHIAN MOGHADDAM, R., KARAHALIOS, K. G., MOGHADDAM, R. Z., AND KARAHALIOS, K. G. Evaluation of Automated Friend Grouping in Online Social Networks. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI EA '14, ACM, pp. 2119–2124.

[49] ETHAN, M. Responsive web design, May 2010. Available at http://alistapart.com/article/responsive-web-design.

[50] FACEBOOK, I. Annual report 2013, Dec. 2013. Available at https://materials.proxyvote.com/Approved/30303M/20140324/AR_200747/.

[51] FAZEEN, M., DANTU, R., AND GUTURU, P. Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches. *Social Network Analysis and Mining 1*, 3 (2011), 241–254.

[52] FERREIRA, A. Effect of online social networking on employee productivity. *SA Journal of Information Management 11*, January (2009), 1–16.

[53] FINN, G. Study: 27% Of Time Online In The US Is Spent On Social Networking, Apr. 2013. Available at http://marketingland.com/study-27-of-time-online-in-the-us-is-spent-on-social-networking-40269.

[54] FIRE, M., PUZIS, R., AND ELOVICI, Y. Organization mining using online social networks. *CoRR abs/1303.3* (2013), 1–19.

[55] GAO, B., BERENDT, B., CLARKE, D., DE WOLF, R., PEETZ, T., PIERSON, J., AND SAYAF, R. Interactive Grouping of Friends in OSN: Towards Online Context Management. *2012 IEEE 12th International Conference on Data Mining Workshops* (Dec. 2012), 555–562.

[56] GAO, H., BARBIER, G., GOOLSBY, R., AND ZENG, D. Harnessing the crowdsourcing power of social media for disaster relief. *ARIZONA STATE UNIV TEMPE* (2011).

[57] GAO, Q., ABEL, F., AND HOUBEN, G.-J. GeniUS: Generic User Modeling Library for the Social Semantic Web. In *Proceedings of the 2011 Joint International Conference on The Semantic Web* (Berlin, Heidelberg, 2012), Springer-Verlag, pp. 160—-175.

[58] GAO, Y., WANG, F., LUAN, H., AND CHUA, T.-S. Brand Data Gathering From Live Social Media Streams. *Proceedings of International Conference on Multimedia Retrieval - ICMR '14* (2014), 169–176.

[59] GARCIA ESPARZA, S., O'MAHONY, M. P. M., SMYTH, B., ESPARZA, S. G., O'MAHONY, M. P. M., AND SMYTH, B. CatStream: Categorising Tweets for User Profiling and Stream Filtering. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (New York, NY, USA, 2013), IUI '13, ACM, pp. 25–36.

[60] GAUCH, S., SPERETTA, M., CHANDRAMOULI, A., AND MICARELLI, A. User Profiles for Personalized Information Access. In *The Adaptive Web*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 54–89.

[61] GHORAB, M. R., ZHOU, D., O?CONNOR, A., AND WADE, V. Personalised Information Retrieval: survey and classification. *User Modeling and User-Adapted Interaction 23*, 4 (May 2013), 381–443.

[62] GILAD, L., ERHARDT, G., MIKE, A., DEVIN, G., IAN, P., AND DANAH, B. The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication 5*, 0 (2011).

[63] GOMEZ-RODRIGUEZ, M., GUMMADI, K. P., AND SCHÖLKOPF, B. Quantifying information overload in social media and its impact on social contagions. *CoRR abs/1403.6838* (2014).

[64] GRAHAM, M., AND AVERY, E. Government Public Relations and Social Media: An Analysis of the Perceptions and Trends of Social Media Use at the Local Government Level. *Public Relations Journal 7*, 4 (2013), 1–21.

[65] HANANI, U., SHAPIRA, B., AND SHOVAL, P. Information Filtering: Overview of Issues, Research and Systems. *User Modeling and User-Adapted Interaction 11*, 3 (Aug. 2001), 203–259.

[66] HANNON, J., BENNETT, M., AND SMYTH, B. Recommending twitter users to follow using content and collaborative filtering approaches. *Proceedings of the fourth ACM conference on Recommender systems - RecSys '10* (2010), 199.

[67] HECHES, Y. Business success and sustainability of major US social network companies. Master's thesis, 2014. MBA in Finance.

[68] HECKMANN, D., SCHWARTZ, T., BRANDHERM, B., SCHMITZ, M., AND VON WILAMOWITZ-MOELLENDORFF, M. Gumo the general user model ontology. In *User Modeling 2005*, L. Ardissono, P. Brna, and A. Mitrovic, Eds., vol. 3538 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, pp. 428–432.

[69] HEIDEMANN, J., KLIER, M., AND PROBST, F. Online social networks: A survey of a global phenomenon. *Computer Networks 56*, 18 (2012), 3866–3878.

[70] HENRI, F., AND PUDELKO, B. Understanding and analysing activity and learning in virtual communities. *Journal of Computer Assisted Learning 19*, 4 (2003), 474–487.

[71] HONG, L., BEKKERMAN, R., ADLER, J., AND DAVISON, B. D. Learning to rank social update streams. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12* (2012), 651.

[72] HOROWITZ, D., AND KAMVAR, S. D. The anatomy of a large-scale social search engine. In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), ACM, pp. 431—-440.

[73] HUGHES, A. L. *Supporting the Social Media Needs of Emergency Public Information Officers with Human-centered Design and Development*. PhD thesis, Boulder, CO, USA, 2012. AAI3549204.

[74] IMRAN, M., CASTILLO, C., DIAZ, F., AND VIEWEG, S. Processing Social Media Messages in Mass Emergency: A Survey. *CoRR abs/1407.7* (2014).

[75] JACOBSON, I. *Object Oriented Software Engineering: A Use Case Driven Approach*, 1 ed. Addison-Wesley Professional, June 1992.

[76] JAVA, A., SONG, X., FININ, T., AND TSENG, B. 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* (San Jose, California, 2007), ACM, pp. 56–65.

[77] JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*, C. Ndellec and C. Rouveirol, Eds., vol. 1398 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 1998, pp. 137–142.

[78] KANE, G. C., ALAVI, M., LABIANCA, G., AND BORGATTI, S. P. What's different about social media networks? A framework and research agenda. *Mis Quarterly 38*, 1 (2014), 275—-304.

[79] KANGAS, P., TOIVONEN, S., AND BÄCK, A., Eds. *" Ads by Google" and Other Social Media Business Models*. Espoo 2007. VTT Tiedotteita ? Research Notes 2384, 2007.

[80] KAPANIPATHI, P., ORLANDI, F., SHETH, A., AND PASSANT, A. Personalized Filtering of the Twitter Stream. *2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation 781* (2011), 1–8.

[81] KAPLAN, A., AND HAENLEIN, M. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons 53*, 1 (Jan. 2010), 59–68.

[82] KHAN, G. F. Social media-based systems: an emerging area of information systems research and practice. *Scientometrics 95*, 1 (Aug. 2012), 159–180.

[83] KIBRIYA, A. M., FRANK, E., PFAHRINGER, B., AND HOLMES, G. Multinomial naive bayes for text categorization revisited. In *Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence* (Berlin, Heidelberg, 2004), AI'04, Springer-Verlag, pp. 488–499.

[84] KIETZMANN, J. H., HERMKENS, K., MCCARTHY, I. P., AND SILVESTRE, B. S. Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons 54*, 3 (May 2011), 241–251.

[85] KIM, W., JEONG, O.-R., AND LEE, S.-W. On social Web sites. *Information Systems 35*, 2 (Apr. 2010), 215–236.

[86] KUSS, D. J., AND GRIFFITHS, M. D. Online social networking and addiction?a review of the psychological literature. *International journal of environmental research and public health 8*, 9 (2011), 3528–3552.

[87] KWAK, H., AND LEE, H. G. A Review of Research on Social Network Services Using the New Media Evolutionary Model. *Informatization Policy 18*, 3 (2011), 3–24.

[88] KWON, O., AND WEN, Y. An empirical study of the factors affecting social network service use. *Computers in Human Behavior 26*, 2 (2010), 254–263.

[89] LAFFERTY, J. How many pages does the average facebook user like?, Apr. 2013. Available at http://allfacebook.com/how-many-pages-does-the-average-facebook-user-like_b115098.

[90] LAI, C. *Semantic indexing modelling of resources in personal and collective memories based on a peer to peer approach.* PhD thesis, University of Compigne, BP 60319 60203 Compigne cedex, july 2011.

[91] LANCASTER, F. W., AND GALLUP, E. Information retrieval on-line. Tech. rep., 1973.

[92] LEWIS, D. Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: ECML-98*, C. Ndellec and C. Rouveirol, Eds., vol. 1398 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 1998, pp. 4–15.

[93] LIETSALA, K., AND SIRKKUNEN, E. *Social media. Introduction to the tools and processes of participatory economy.* Tampere University Press, 2008.

[94] LIM, K. H., AND DATTA, A. Finding Twitter Communities with Common Interests Using Following Links of Celebrities. In *Proceedings of the 3rd International Workshop on Modeling Social Media* (New York, NY, USA, 2012), MSM '12, ACM, pp. 25–32.

[95] LINKEDIN. Linkedin reaches 300 million members worldwide, Apr. 2014. Available at http://press.linkedin.com/News-Releases/333/LinkedIn-reaches-300-million-members-worldwide.

[96] LUCHMAN, J. N., BERGSTROM, J., AND KRULIKOWSKI, C. A motives framework of social media website use: A survey of young Americans. *Computers in Human Behavior 38* (Sept. 2014), 136–141.

[97] MAINKA, A., HARTMANN, S., STOCK, W., AND PETERS, I. Government and Social Media: A Case Study of 31 Informational World Cities. In *Proceedings of the 47th Hawaii International Conference on System Sciences* (Hawaii, 2014), IEEE Computer Society, pp. 1715–1724.

[98] MANGOLD, W. G., AND FAULDS, D. J. Social media: The new hybrid element of the promotion mix. *Business Horizons 52*, 4 (2009), 357 – 365.

[99] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA, 2008.

[100] MARSDEN, P. V., AND FRIEDKIN, N. E. Network Studies of Social Influence. *Sociological Methods & Research 22*, 1 (1993), 127–151.

[101] MAY, A., CHAINTREAU, A., KORULA, N., AND LATTANZI, S. Filter & follow: How social media foster content curation. In *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems* (New York, NY, USA, 2014), SIGMETRICS '14, ACM, pp. 43–55.

[102] MAYFIELD, A. *What is social media ?* 2008.

[103] MCCANDLESS, M., HATCHER, E., AND GOSPODNETIC, O. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0.* Manning Publications Co., Greenwich, CT, USA, 2010.

[104] MCPHERSON, M., SMITH-LOVIN, L., AND COOK, J. M. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology 27*, 1 (2001), 415–444.

[105] MENDES, P. P. N., PASSANT, A., AND KAPANIPATHI, P. Twarql: Tapping into the Wisdom of the Crowd. In *Proceedings of the 6th International Conference on Semantic Systems* (New York, NY, USA, 2010), I-SEMANTICS '10, ACM, pp. 45:1—-45:3.

[106] MEO, P. D., NOCERA, A., TERRACINA, G., AND URSINO, D. Recommendation of similar users, resources and social networks in a Social Internetworking Scenario. *Information Sciences 181*, 7 (Apr. 2011), 1285–1305.

[107] MERGEL, I. *Social Media in the Public Sector: A Guide to Participation, Collaboration and Transparency in The Networked World.* Jossey-Bass, 2012.

[108] MOSTAFA, M. M. More than words: Social networks? text mining for consumer brand sentiments. *Expert Systems with Applications 40*, 10 (2013), 4241–4251.

[109] MURTAGH, R. The role of #hashtags in social media and search, Nov. 2013.

[110] NAAMAN, M., BOASE, J., AND LAI, C.-H. Is it really about me ? message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (2010), pp. 189–192.

[111] O'BANION, S., BIRNBAUM, L., HAMMOND, K., AND BANION, S. O. Social media-driven news personalization. *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web - RSWeb '12* (2012), 45.

[112] O'REILLY, T. What is web 2.0? design patterns and business models for the next generation of software.

[113] ORLANDI, F., BRESLIN, J., AND PASSANT, A. Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th International Conference on Semantic Systems* (New York, NY, USA, 2012), ACM, pp. 41–48.

[114] OSTER, G., MOLLI, P., DUMITRIU, S., AND MONDJAR, R. Uniwiki: A collaborative p2p system for distributed wiki applications. In *in "18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises - WETICE 2009, Pays-Bas Groningen", IEEE Computer Society* (2009), pp. 87–92.

[115] PASSANT, A., HASTRUP, T., AND BOJ, U. Microblogging : A Semantic and Distributed Approach. In *4th Workshop on Scripting for the Semantic Web (SFSW 2008), in conjunction with ESWC 2008* (2008), pp. 1–12.

[116] PHELAN, O., MCCARTHY, K., AND SMYTH, B. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems* (2009), ACM, pp. 385–388.

[117] PHUVIPADAWAT, S., AND MURATA, T. Breaking news detection and tracking in twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Aug 2010), vol. 3, pp. 120–123.

[118] PLUMBAUM, T., SCHULZ, K., KURZE, M., AND ALBAYRAK, S. My Personal User Interface: A Semantic User-Centric Approach to Manage and Share User Information. In *Human Interface and the Management of Information. Interacting with Information*, M. Smith and G. Salvendy, Eds., vol. 6771 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011, pp. 585–593.

[119] PLUMBAUM, T., WU, S., DE LUCA., E. W., AND ALBAYRAK, S. User Modeling for the Social Semantic Web. In *In Proceedings of iswc2011* (2011), pp. 78–89.

[120] QU, Z., AND LIU, Y. Interactive Group Suggesting for Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2* (Stroudsburg, PA, USA, 2011), HLT '11, Association for Computational Linguistics, pp. 519–523.

[121] RAKESH, V., SINGH, D., VINZAMURI, B., AND REDDY, C. C. K. Personalized Recommendation of Twitter Lists Using Content and Network Information. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (Michigan, USA, 2014), The AAAI Press, pp. 416–425.

[122] RIJSBERGEN, C. J. V. *Information Retrieval*, 2nd ed. Butterworth-Heinemann, Newton, MA, USA, 1979.

[123] RIZZO, G., AND TRONCY, R. {NERD}: evaluating named entity recognition tools in the web of data. In {ISWC} 2011, {W}orkshop on {W}eb {S}cale {K}nowledge {E}xtraction ({WEKEX}'11), {O}ctober 23-27, 2011, {B}onn, {G}ermany ({B}onn, {GERMANY}, 2011).

[124] ROWE, M., AND CIRAVEGNA, F. Getting to me–exporting semantic social network information from facebook. In *The 7th International Semantic Web Conference* (2008), Citeseer, p. 43.

[125] RUSSELL, M., AND RUSSELL, M. *Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*. Head First Series. O'Reilly Media, Incorporated, 2011.

[126] SAKAKI, T., OKAZAKI, M., AND MATSUO, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 851–860.

[127] SALTON, G., FOX, E. A., AND WU, H. Extended boolean information retrieval. *Commun. ACM 26*, 11 (Nov. 1983), 1022–1036.

[128] SALTON, G., WONG, A., AND YANG, C. S. A vector space model for automatic indexing. *Commun. ACM 18*, 11 (Nov. 1975), 613–620.

[129] SANKARANARAYANAN, J., SAMET, H., TEITLER, B. E., LIEBERMAN, M. D., AND SPERLING, J. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2009), GIS '09, ACM, pp. 42–51.

[130] SATO, G. Y., AND BARTHÈS, J.-P. A. Copboard: A catalyst for distributed communities of practice. *International Journal of Software Science and Computational Intelligence (IJSSCI) 2*, 1 (2010), 52–71.

[131] SCHOLLMEIER, R. A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications. In *Proceedings of the First International Conference on Peer-to-Peer Computing* (Washington, DC, USA, 2001), IEEE Computer Society, pp. 101–102.

[132] SEBASTIANI, F. Text categorization. In *Text Mining and its Applications to Intelligence, CRM and Knowledge Management* (2005), WIT Press, pp. 109–129.

[133] SEONG, S.-W., SEO, J., NASIELSKI, M., SENGUPTA, D., HANGAL, S., TEH, S. K., CHU, R., DODSON, B., AND LAM, M. S. Prpl: A decentralized social networking infrastructure. In *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond* (New York, NY, USA, 2010), MCS '10, ACM, pp. 8:1–8:8.

[134] SHAPIRA, B., ROKACH, L., AND FREILIKHMAN, S. Facebook single and cross domain data for recommendation systems. *User Modeling and User-Adapted Interaction 23*, 2-3 (Sept. 2012), 211–247.

[135] SHEA, B. Social media stats 2014 [infographic], Jan. 2014. Available at `http://www.mediabistro.com/alltwitter/social-media-stats-2014_b54243`.

[136] SHEN, K., WU, J., ZHANG, Y., HAN, Y., YANG, X., SONG, L., AND GU, X. Reorder user's tweets. *ACM Transactions on Intelligent Systems and Technology 4*, 1 (Jan. 2013), 1–17.

[137] SHIH, C. *The Facebook Era: Tapping Online Social Networks to Build Better Products, Reach New Audiences, and Sell More Stuff*. Pearson Education, 2009.

[138] SMITH, A. The internet's role in campaign 2008. *Pew Internet & American Life Project 15* (2009).

[139] SNELL, J., ATKINS, M., NORRIS, W., MESSINA, C., WILKINSON, M., AND DOLIN, R. Json activity streams 1.0. *Abgerufen am 22*, 08 (2011), 2013.

[140] SOREN, A., LEE, F., DANIEL, M., ANGELA, F., AND JUAN, S. Use cases and requirements for mapping relational databases to rdf, June 2010. Available at `http://www.w3.org/TR/rdb2rdf-ucr/`.

[141] SPEIER, C., VALACICH, J. S., AND VESSEY, I. The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences 30*, 2 (1999), 337–360.

[142] SPENCE, R., AND PRESS, A. *Information Visualization*. Addison Wesley, Dec. 2000.

[143] SRIRAM, B., FUHRY, D., DEMIR, E., FERHATOSMANOGLU, H., AND DEMIRBAS, M. 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* (New York, USA, 2010), ACM, pp. 841–842.

[144] SUN, Y., AND SHANG, R.-A. The interplay between users? intraorganizational social media use and social capital. *Computers in Human Behavior* (June 2014).

[145] SUTTON, J., PALEN, L., AND SHKLOVSKI, I. Backchannels on the front lines: Emergent uses of social media in the 2007 southern California wildfires. In *Proceedings of the 5th International ISCRAM Conference* (Washington, DC, 2008), no. May, pp. 624–632.

[146] TIM, B.-L. Linked data - design issues, July 2006. Available at `http://www.w3.org/DesignIssues/LinkedData.html`.

[147] TRAMP, S., FRISCHMUTH, P., AND ERMILOV, T. An architecture of a distributed semantic social network. *Semantic Web 5*, 1 (2014), 77–95.

[148] TWITTER. About twitter, inc., 2014. Available at https://about.twitter.com/company.

[149] VIRMANI, C., PILLAI, A., AND JUNEJA, D. Study and analysis of Social network Aggregator. In *Optimization, Reliabilty, and Information Technology (ICROIT)* (India, 2014), pp. 145–148.

[150] WENG, J., LIM, E.-P., JIANG, J., AND HE, Q. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2010), WSDM '10, ACM, pp. 261–270.

[151] WIDMAN, J. Edgerank, a guide to facebook's newfeed algorithm. Available at http://edgerank.net/.

[152] WIKIPEDIA. Wikipedia:multilingual statistics, 2014. Available at https://about.twitter.com/company.

[153] WILKINSON, D., AND THELWALL, M. Researching Personal Information on the Public Web: Methods and Ethics. *Social Science Computer Review 29*, 4 (Aug. 2010), 387–401.

[154] WU, M. Community vs. social network, july 2010. Available at http://lithosphere.lithium.com/t5/science-of-social-blog/Community-vs-Social-Network/ba-p/5283.

[155] YEUNG, C.-M. A., LICCARDI, I., LU, K., SENEVIRATNE, O., AND BERNERS-LEE, T. Decentralization: The future of online social networking. In *W3C Workshop on the Future of Social Networking Position Papers* (2009), vol. 2, pp. 2–7.

[156] YOUNG, A. L., AND QUAN-HAASE, A. Information revelation and internet privacy concerns on social network sites: A case study of facebook. In *Proceedings of the Fourth International Conference on Communities and Technologies* (New York, NY, USA, 2009), C&T '09, ACM, pp. 265–274.

[157] YOUNG, S. 28 must see social media statistics, Aug. 2013. Available at http://www.socialmediatoday.com/content/28-must-see-social-media-statistics.

[158] YOUSSEF, B. E. Online Social Network Internetworking Analysis. *International Journal of Next-Generation Networks (IJNGN) 6*, 2 (2014), 1–15.

[159] ZELANDIYA. So, you need to understand language data? open-source nlp software can help!, july 2014. Available at http://entopix.com/so-you-need-to-understand-language-data-open-source-nlp-software-can-help/.

[160] Zhang, J., Wang, Y., and Vassileva, J. SocConnect: A Personalized Social Network Aggregator and Recommender. *Inf. Process. Manage. 49*, May, 2013 (2012), 721—-737.

[161] Zhang, Y., and Leung, L. A review of social networking service (SNS) research in communication journals from 2006 to 2011. *New Media & Society* (Jan. 2014), 1–18.