UNIVERSIDADE DE VIGO

ESCOLA DE
ENXEÑARÍA DE TELECOMUNICACIÓN

Ph.D. programme in Telematics Engineering

Ph.D. Thesis
Submitted for the International Doctor mention

Social Data Mining Strategies for User Modelling with Personalisation Purposes

Author: Sandra Servia-Rodríguez
Advisors: Ana Fernández-Vilas
Rebeca P. Díaz-Redondo

2015

Abstract

The abundance of information in the online world results in a growing demand for relevant content, making any service in this medium a perfect environment within which personalisation could blossom. The availability of information about users' interests, opinions and so on in online social sites facilitates the effective modelling of users and services, enabling avoidance of the well-known issues in personalisation that come up when new users or items are added to the system –the "cold-start problem"–. In an attempt to externalise the provision of personalisation to online services, thereby allowing them to be exclusively focused on their tasks, the thesis of this dissertation is that an intermediary model of the user constructed by properly mining user generated content in social media can be exploited to create or improve technological social applications. The model we propose is based on representing users' online life by means of what we call *social spheres* and considering only users' data available from public APIs (private messages, retweets, etc.) with users' permission. Two key contributions of this model are (i) a methodology to extract the thematic fields users talk about with their social media contacts and (ii) a measure of the strength of the tie between two individuals from their interaction data available in social media sites. For the former, we use several data mining techniques to represent users' interests or *social contexts* by means of tags of representative words and validate this proposal by using Twitter data. We also show how these social contexts could be used to improve an important marketing application, namely that of advertising recommendation. For the latter, and contrary to previous approaches, we take into account different interaction types and contexts, the time in which interactions occur, the people involved in them and the frequency of interactions with the rest

of the user's contacts, finding that our measure assesses with high accuracy users' perceived strength of their social ties. We finally discuss how this model of social spheres may be exploited to improve a wide range of technological applications, from recommender systems to e-mail readers, and describe two of them in detail: an application that helps users gain attention in social media and other designed to find trustworthy users in these media. We also present a prototype of an intermediary service that obtains these social spheres and makes them available to other services.

Acknowledgments

En primer lugar quiero expresar mi más sincero agradecimiento a mis tutoras, Ana Fernández Vilas y Rebeca P. Díaz Redondo, por haberme dado la oportunidad de realizar esta tesis doctoral y haberme guiado en el largo y duro camino hacia ella. Gracias por la confianza depositada en mí y por su ayuda y esfuerzo incondicional para que esta tesis fuese una realidad.

I would also like to extend my sincere gratitude to Cecilia Mascolo and Bernardo Huberman who guided me during my stays at the University of Cambridge and the HP Labs respectively. I thank Cecilia for teaching me the importance of collaborations and perfection in research; and Bernardo for showing me that there is a lot of interesting research beyond academia. Thanks for your guidance, for those interesting conversations and motivating words.

I am grateful to those that I met during my stays at Cambridge and Palo Alto, and specially to those with whom I have co-authored papers: Anastasios Noulas and Sitaram Asur. A special gratitude to Deborah Falcone, thanks to whom I had a wonderful time during my stay in Cambridge; and to Chloë Brown for her assistance and helpful comments.

Gracias a mis amigos de los laboratorios de Inteligencia, TSC-5 y Gradiant, con quienes he compartido infinidad de comidas, cafés y alguna que otra cena. Gracias por haber hecho mis días en el CUVI mucho más llevaderos. Agradezco también por ello a mis compañeros del laboratorio B-004, presentes y pasados.

Agradecer al Ministerio de Economía y Competitividad (antiguo Ministerio de Ciencia e Innovación) el haberme otorgado una Ayuda FPI (BES-2011-046878)

para sufragar mis estudios de doctorado, y a los miembros del Grupo de Servicios de la Sociedad de la Información por confiar en mí para formar parte del proyecto CLOUDIA asociado a dicha ayuda.

Por último, y más importante, agradecer a mis padres, a mis hermanos, y a mi abuela por el apoyo brindado durante estos años, sin el cual esta tesis no habría sido ni siquiera empezada. Y a mis amigos, Antía, Minia, Pablo, Paula, Rocío y Tania, por escuchar mis incesantes quejas y ayudarme cada día a ver el vaso medio lleno.

A todos los demás que no he nombrado y que me habéis acompañado y animado a lo largo de esta etapa, gracias.

Contents

1

Introduction

The advent of the World Wide Web in the early 1990s has changed the way in which modern societies relate, overcoming the constraints imposed by the physical world and allowing people to communicate with others thousands of miles away. Initially intended to make it easier for nuclear physics researchers to share information, it has evolved to become a medium that the public use to communicate, find information, and even for entertainment. Although since its early days there have been initiatives to make the Web more social, it was the emergence of Social Web technologies in the early 2000s that provoked the socialisation of the Web through the active participation and involvement of the users, since they started to act not only as typical consumers, but as producers of information. These technologies have provided individuals with powerful tools to freely disseminate factual information, opinions, and, ultimately, any sort of content that they wish to share with their social circles, developing new methods of communication that go beyond traditional face-to-face interactions. Within this general definition, there are various types of social media: blogs (Blogger, WordPress, ...), social

networking sites (Facebook, Foursquare, ...), collaborative projects (Wikipedia, OpenStreetMap, ...), content communities (YouTube, LastFM, ...), etc. Although there is not a systematic way to categorise social media applications, some researchers such as Kaplan and Haenlein [KH10] proposed a categorisation based on different theories of media research and social processes. In addition, even within each type of social media, there are differences between the features that sites provide to their users to share content: the type of content to share (text, photos, videos, ...), visibility of the shared content (everyone, just *friends*, ...), purpose and scope of the site (personal, business-oriented, ...), etc. Regardless of the specific social media technology considered, the content of these shared items, together with other metadata (time, location, ...), makes them valuable data sources that reflect users' interests, users' opinions and so on [GL12]. The pervasive use of these technologies (as of September 2014, Facebook, the most popular social networking site, had more than one billion monthly active users [Face]) has entailed the availability of large amounts of this dynamic and continuously updated user generated content whose analysis has applications across several domains, from business to social sciences [BL11, KH10, GK09, BGL10].

Personalisation, the use of technology to accommodate the differences between individuals, has played an important role in the success of online services. For years, personalisation has involved applications collecting user information, which, after appropriate analysis, they used to deliver appropriate content. User information was traditionally obtained from a history of previous sessions, or through interactions in real time [Bon01]. This approach has various disadvantages and limitations mainly related to (i) the absence of information when a new service is delivered or a new user uses the service for first time –*the cold start problem* [SPUP02]– and (ii) that even the most active users only have rated a small subset of the available services, which makes the data sparse and insufficient to identify similarities in users' interests –*the sparsity problem* [SKKR01]. More recently, the emergence of social media technologies and their enormous popularity have transformed the Web into an universe swarming with user generated content. The great acceptance of these technologies, their penetration in all social sectors and the users' freedom to participate suggest that the use of this user generated content for personalisation purposes would greatly benefit services [BB07, Bir07], even allowing the cold-start and the sparsity problems to

be overcome. In their aim of satisfying their users, services need to be aware of their interests, what they like. Most social media technologies allow users to create profiles including demographic and geographic data and even interests, and the mere analysis of this information is enough to develop powerful personalised services, such as the advertising services on Facebook [Facb] and Youtube [You]. Researchers have also proven the usefulness of mining the spontaneous content that users unconsciously post in their accounts to create compelling experiences that encourage users to keep using the services [TDH05, JWL+11, AGHT11].

In addition, these Social Web technologies, and especially online social networking services, have enabled people to connect with one another, forging and strengthening online relationships. This, together with the tendency of individuals to associate and bond with similar others (*homophily*) observed by sociologists [MSLC01], suggests the benefits of enhancing service personalisation further by considering the interests of users' friends. Although many social media technologies, and particularly online social networks, allow individuals to connect with others, not all these connections are indicative of actual relationships. On the contrary, a person can only maintain a limited number of social relationships, and even a limited number of them at different levels of closeness [Dun98]. It is then necessary to separate the sheep from the goats to develop effective *socially-enhanced services*.

Nearly at the same pace as the above social trend, another revolutionary phenomenon has rushed into the technological landscape: the so-called service-oriented computing (SOC) paradigm [Pap03]. This computing paradigm has changed the way software applications are designed, delivered and consumed. In SOC, services are used as basic blocks to construct rapid, low-cost, secure and reliable applications, reducing the need to develop new software components each time a new business process arises [PVDH07, PTDL08]. By using standard description languages a service can expose its interface to the outside world for discovery and being invoked separately or as a composition of multiple services. As outstanding examples, companies such as Google, Amazon, Twitter and Facebook have offered Web services to provide access to some of their resources, enabling third parties to combine and reuse their services [SQV+14]. The unavoidable penetration of SOC and its strongly-related Cloud Computing

paradigm [WB10, Rai09] – whose so-called Everything as a Service enables a Cloud (a metaphor for the Internet) that hosts resources that will be delivered as services at a high level of granularity and that may be composed in a flexible manner in response to complex necessities [ZCB10] – suggests the need for applications to be developed in such a way that they can be integrated into existing services or/and built upon them.

The previous suggests that personalisation of online services should not stand on the sidelines of this trend. That is, the provision of personalisation should be delivered as intermediary services to be discovered and invoked by other online services that wish to offer personalised experiences to their customers, releasing them from discovering their customers' preferences. Also, services in charge of delivering users' preferences – users' profiles – should not dismiss the potential of the user generated content available in social media for obtaining users' preferences and any other useful information. Besides, individuals use several social media sites at a time, and the integration of the content that they produce in all these sites is what entirely characterises their online life. To develop such services, several issues need to be solved: how to properly represent/model users, how to properly mine user generated content to extract useful information, how to distinguish users' true friends from simple acquaintances, etc. In this dissertation, we focus on these questions and, specifically, on proposing and analysing different data mining techniques to extract useful information from user generated content in different social media sites and represent it in such a way that it can be properly delivered to other services, in order for them to be successfully personalised or socially-enhanced.

1.1. Thesis and its substantiation

The thesis explored in this dissertation is that *an intermediary model of the user constructed by properly mining user generated content in different social media sites can be exploited to create or improve technological social applications.* In order to examine this thesis, it is necessary to design such a model in a way that encompasses users' interests and contacts, and to consider how this model might be used by such technological applications.

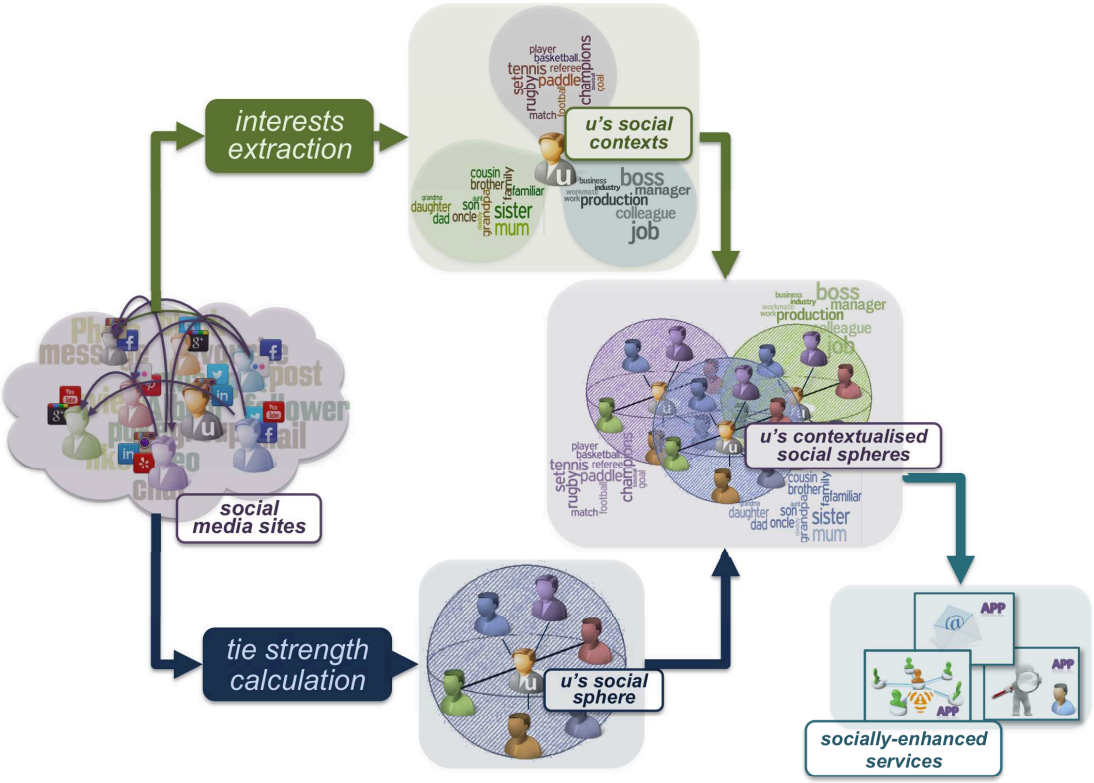


Figure 1.1: A model of social spheres

More specifically, in this dissertation we propose a user-centred model for personalising applications based on mining data from different social web sites to identify the subjects – topics – of interests of users and the strength of the relationships between them and all the users with whom they interact. Figure 1.1 provides an overview of the different outputs provided by this model: *social contexts*, *social spheres* and *contextualised social spheres* of each user. With *social contexts* we refer to the topics of interests of the user, extracted from mining textual content posted by the user (and usually shared with their contacts) and delivered in the form of tag clouds of representative words. *Social spheres* is the term we coined to refer to the set of users with whom the target user usually interacts through social media sites, together with the strength of their tie inferred from their interactions in these platforms. Finally, when merging contexts and contacts, the *contextualised social spheres* emerge, being each sphere made of those users with whom the target user frequently talks about the topic – context – of the sphere, together with the strength of their ties taking into account only

those interactions in the scope of the topic. In the rest of this dissertation, we detail the techniques that we propose to compute these contexts and spheres. We also describe how to successfully implement this model in a prototype of service in charge of both monitoring users' activity in social media sites and providing the contexts and spheres to personalise/socially-enhance a wide range of applications and services, also described. Given the sensitivity of the managed information and the increasing concern of keeping this kind of information private, such a service should require users' permission to both access their social data on their behalf and provide applications with their social contexts and/or spheres.

1.2. Contributions and outline

1.2.1. Contributions

The main contribution of this research is the externalisation of the provision of personalisation to online services through the definition of a *social spheres* model using interaction data retrieved from several social media sites. In order to create these social spheres, we also propose strategies:

1. To discover and model the interests of social media individuals and those contacts that share these interests with them through the analysis of the available abundance of user generated content in social media sites; and to represent these *social contexts* by means of tags of representative words that simplify their use by almost every application,

2. To measure and represent the strength of the tie between two social media individuals from the perspective of one of them through the analysis of signs of interaction in social sites available from their *Application Programming Interfaces* – APIs – (private messages, retweets, mentions, etc.) with their permission; and contrary to previous approaches, taking into account different types of interaction and contexts, the time in which interactions occur, the people involved in them and the frequency of interactions with the rest of the user's contacts, and

3. To show how this social spheres model can be used to create and enhance existing social technological services such as applications to gain attention or find trustworthy experts in social media, and how it can be easily integrated into a SOC-service that delivers these spheres to other services on request and always with users' permission.

1.2.2. Outline

During these PhD studies I have been involved in many fruitful collaborations that have yielded several published works in various peer-reviewed journals and conferences that span the areas of social computing, social networks, topic modelling and ambient intelligence. More specifically, the rest of this dissertation is structured as follows, indicating in brackets the publications they are based on.

In **Chapter** 2 we outline the existing work in this area and describe how our research relates to and builds upon this.

In **Chapter 3**, we present our methodology to extract the thematic fields users talk about with their social media contacts, i.e. their social contexts. Using several data mining techniques, we are able to represent each context by means of tags of representative words and validate this proposal using Twitter data. We then examine how social contexts could be used to improve an important marketing application on Facebook, namely that of advertisement recommendation ([SFDP13b, SFDP13a]).

[**SFDP13b**] Sandra Servia-Rodríguez, Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and José J. Pazos-Arias. Inferring Contexts from Facebook Interactions: A Social Publicity Scenario. *IEEE Transactions on Multimedia*, 15(6):1296–1303, October 2013.

[**SFDP13a**] Sandra Servia-Rodríguez, Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and José J. Pazos-Arias. Comparing Tag Clustering Algorithms for Mining Twitter Users' Interests. In *International Conference on Social Computing (SocialCom)*, pages 679–684, Washington D.C., USA, September 2013.

In **Chapter 4**, we study how to measure the strength of users' ties by using signs of interaction available from social sites' APIs (private messages, retweets, mentions, etc.) with users' permission. To this aim, and contrary to previous

approaches, we take into account (i) different interaction types and contexts, (ii) the time in which interactions occur, (iii) the people involved in them and (iv) the frequency of interaction with the rest of the user's contacts. Through a user study on Facebook, we find that our model represents with high accuracy users' perceived strength of their social ties ([SDF+14, FDS14, SDFP13, SFDP12, SDFP12]).

[SDF+14] Sandra Servia-Rodríguez, Rebeca P. Díaz-Redondo, Ana Fernández-Vilas, Yolanda Blanco-Fernández, and José J. Pazos-Arias. A tie strength based model to socially-enhance applications and its enabling implementation: mySocialSphere. *Expert Systems with Applications*, 41(5):2582 – 2594, 2014.

[FDS14] Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and Sandra Servia-Rodríguez. IPTV parental control: A collaborative model for the Social Web. *Information Systems Frontiers*, pages 1–16, 2014.

[SDFP13] Sandra Servia-Rodríguez, Rebeca P. Díaz-Redondo, Ana Fernández-Vilas, and José J. Pazos-Arias. Mining Facebook Activity to Discover Social Ties: Towards a Social-Sensitive Ecosystem. In Ivan I. Ivanov, Marten van Sinderen, Frank Leymann, and Tony Shan, editors, *Cloud Computing and Services Science*, volume 367 of *Communications in Computer and Information Science*, pages 71–85. Springer International Publishing, 2013.

[SDFP12] Sandra Servia-Rodríguez, Rebeca Díaz-Redondo, Ana Fernández-Vilas, and J Pazos-Arias. Using Facebook activity to infer social ties. In *International Conference on Cloud Computing and Services Science, CLOSER*, Porto, Portugal, April 2012.

[SFDP12] Sandra Servia-Rodríguez, Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and José J. Pazos-Arias. Inferring Ties for Social-Aware Ambient Intelligence: The Facebook Case. In *International Symposium on Ambient Intelligent (ISAMI)*, volume 153 of *Advances in Intelligent and Soft Computing*, pages 75–83, Salamanca, Spain, March 2012. Springer Berlin Heidelberg.

In **Chapter 5**, we discuss how our model of social spheres may be exploited to improve a wide range of technological applications, from recommender systems to e-mail readers, and present two of them in detail: an application that helps users gain attention in social media and other designed to find trustworthy users in these media. We also present a prototype of a service that obtains these social spheres and makes them available to other services ([SDF+14, SDF15, DFPS12]).

[SDF+14] Sandra Servia-Rodríguez, Rebeca P. Díaz-Redondo, Ana Fernández-Vilas, Yolanda Blanco-Fernández, and José J. Pazos-Arias. A tie strength based model to socially-enhance applications and its enabling implementation: mySocialSphere. *Expert Systems with Applications*, 41(5):2582 – 2594, 2014.

[SDF15] Sandra Servia-Rodríguez, Rebeca P. Díaz-Redondo, and Ana Fernández-Vilas. Are tweets biased by audience? an analysis from the view of topic diversity. In *International Social Computing, Behavioral-Cultural Modeling and Prediction Conference (SBP'15)*, Lecture Notes in Computer Science, Washington D.C., USA, April 2015.

[DFPS12] Rebeca P. Díaz-Redondo, Ana Fernández-Vilas, José J. Pazos-Arias, and Sandra Servia-Rodríguez. A Social P2P Approach for Personal Knowledge Management in the Cloud. In *On the Move to Meaningful Internet Systems: OTM 2012 Workshops*, volume 7567 of *Lecture Notes in Computer Science*, pages 585–594, Rome, Italy, September 2012. Springer Berlin Heidelberg.

In **Chapter 6**, we review the contributions of our research and draw conclusions, as well as identifying directions for future research that may be built on the work described on this dissertation.

Regarding structure, with the exception of Chapter 2 and Chapter 6, each chapter begins with a general introduction to the addressed topic that serves both to indicate the contributions of the chapter and to relate these contributions with the thesis explored in this dissertation and with previous chapters (if applicable). Then, the models, analysis and results obtained in the scope of the contributions are detailed. Each chapter ends with two sections: *Discussion* and *Summary* respectively. The former serves to state our interpretations and opinions as well as to explain the implications of our findings, while the latter summarises our contributions and findings, and relates these findings with the thesis that we explore.

1.2.3. Other works during PhD study

Apart from the published works on which this dissertation is based, I have been involved in other research work that have led me either to tackle the thesis of this dissertation or be aware of further research in the area. Specifically, [BLB+12] and [MGR+12] belong to the former group, whereas the latter is composed of [SNM+15, SHA15, SNM+14] and [BDS+14]. We will come back to the latter in Section 6.2.

[SNM+15] Sandra Servia-Rodríguez, Anastasios Noulas, Cecilia Mascolo, Ana Fernández-Vilas, and Rebeca P. Díaz-Redondo. The evolution of your success lies at the centre of your co-authorship network. *PLoS ONE*, 10:e0114302, 03 2015.

[SHA15] Sandra Servia-Rodríguez, Bernardo A. Huberman, and Sitaram Asur. Deciding what to display: maximizing the information value of social media. In *Workshop on Modeling and Mining Temporal Interactions (M2TI) at ICWSM' 15*, Oxford, UK, May 2015.

[SNM+14] Sandra Servia-Rodríguez, Anastasios Noulas, Cecilia Mascolo, Ana Fernández-Vilas, and Rebeca P. Díaz-Redondo. The evolution of your success lies in the centre of your co-authorship network. In *Quantifying Success (2.0) –co-located with ECCS 2014*, Lucca, Italy, September 2014.

[BDS+14] Mohamed Ben-Khalifa, Rebeca P. Díaz-Redondo, Sandra Servia-Rodríguez, Ana Fernández-Vilas, and Rafael López-Serrano. Is There a Crowd? Experiences in using Density-Based Clustering and Outlier Detection. In *International Conference on Mining Intelligence and Knowledge Exploration (MIKE)*, Cork, Ireland, December 2014.

[BLB+12] Jack F. Bravo-Torres, Martín López-Nores, Yolanda Blanco-Fernández, Sandra Servia-Rodríguez, and Jorge García-Duque. A virtualization layer for mobile consumer devices to support demanding communication services in vehicular ad-hoc networks. In *IEEE International Conference on Consumer Electronics (ICCE)*, pages 225–226, Las Vegas, USA, January 2012.

[MGR+12] Manuela I. Martín-Vicente, Alberto Gil-Solla, Manuel Ramos-Cabrer, Yolanda Blanco-Fernández, and Sandra Servia-Rodríguez. Semantics-driven recommendation of coupons through Digital TV: Exploiting synergies with social networks. In *IEEE International Conference on Consumer Electronics (ICCE)*, pages 564–565, Las Vegas, USA, January 2012.

2

User modelling and personalisation in social media

Social media technologies have entailed the availability of large amounts of data about individuals whose analysis and application to successfully personalise online services have been studied for a long time [GZR+10, RSvZ10, HMI03]. As we have seen in the previous chapter, most of this personalisation uses data from static user profiles, which can include demographic and geographic data, social contacts and even interests [Facb, You]. But apart from creating a profile, social media platforms usually allow individuals to post content and interact with others, freely manifesting their interests and forging social ties. Although the study and exploitation of this user generated content to create or improve technological social applications have been topics of interest for many researchers [AGHT11, BSH+10, KH10, GK09, NSV11], most previous work has focused on mining content from only one social media site, and on personalising/socially-enhancing applications under the umbrella of the given

site [RSvZ10]. Little research has specifically studied how to mine and integrate users' data from different social media platforms to create models of the user – in terms of social contacts and interests – that allow the personalisation of applications or services independently of the platform that hosts their data.

2.1. Extraction of interests

2.1.1. Traditional text mining approaches

Text mining, roughly equivalent to text analytics, is an active area of research in Computer Science that tries to automatically derive high-quality information from text. It provides a solution for the crisis of information overload based on combining techniques from data mining, machine learning, natural language processing, information retrieval and knowledge management [FS07], and its application spans many domains, from the World Wide Web to Business Intelligence. This vague definition of text mining broadly encompasses several related topics and algorithms for text analysis, such as feature extraction, which aims to identify entities and their relations in the text, or text summarisation, whose goal is to reduce the amount of text in a document while still keeping its key meaning.

Regardless of the applied technique and the possible applications, the main goal in text mining is the effective representation of the content of text documents. Traditionally, a document is represented as a bag-of-words, assuming that the words occur independently in the document. This results in a vector representation, where bag-of-words vectors have a very high dimensionality – each dimension corresponding to one term from the language. In order to analyse the concepts in the documents, a lower-dimensional semantic space is desired. To obtain this reduction, researchers have proposed *clustering* documents by segmenting a corpus of documents into partitions, each corresponding to a topical cluster [vR79, KS97]. In this case, the dimensional reduction comes from viewing each cluster as a dimension. In contrast, *dimensional reduction* looks for a lower-dimensional representation that is faithful to the original representation. *Topic modelling* arises when integrating soft clustering (clustering performed in such a

way that each document belongs to different clusters with different membership probabilities) with dimension reduction. That is, when clustering is performed by a probabilistic model where each document belongs to different clusters with different membership probabilities.

The well-known Latent Semantic Indexing (LSI) technique was introduced by Deerwester et al. in 1990 [DDL+90]. This automatic indexing method projects both documents and terms into a low dimensional space that represents their semantic concepts in the document using, to this aim, the singular value decomposition (SVD) of the term-document matrix. Later, Hoffman [Hof99] proposed the probabilistic Latent Semantic Indexing (pLSI) technique by extending LSI in a probabilistic context. This approach uses a latent variable model that represents documents as mixtures of topics, outperforming LSI in a vector space model framework. However, pLSI contains a large number of parameters that grows linearly with the number of documents, besides there is no natural way to compute the probability of an arbitrary document not in the training data. Being aware of these limitations, Blei et al. [BNJ03] proposed Latent Dirichlet Allocation (LDA), a probabilistic technique that includes a process for generating the topics in each document, reducing the number of parameters to be learned and providing a clearly-defined probability for arbitrary documents. LDA has quickly become one of the most popular probabilistic text modelling techniques in machine learning, being shown to be effective in some text-related tasks such as document classification.

2.1.2. Extraction of interests in social media

Textual data in social media presents new challenges and opportunities due to its characteristics differing from traditional textual data [AZ12]. User generated content in social media is time-dependent, which means that the text in social media is not independent and identically distributed data, but users post comments about recent events, such as new products, movies, sports, etc. and, sometimes, influenced by their contacts. In addition, some social media sites restrict the length of user-created content such as mocroblogging messages, QA passages, product reviews, etc. This is the case for Twitter, which limits the

size of each tweet to 140 characters or Picasa, which limits comments to 512 characters. Therefore, text analysis methods need to be capable of successfully processing short texts, which usually consist of a few phrases or sentences with sparse contextual information. This makes it very difficult for bag-of-words-based models to build semantic connections between words. Solutions proposed so far have taken advantage of external sources to fix the semantic gap and find given connections [SP06, CV07, GM07].

User generated content usually presents high variance in the quality of the content. This happens mainly because some users are experts on the topic and post information very carefully, while others post the first idea that comes to their minds. Users also use new abbreviations and acronyms that seldom appear in conventional text documents, which makes it difficult to identify their semantic meaning in the text. Apart from the content itself, there is meta-content available in social media, such as *hashtags* in Twitter (keywords identified with the symbol "#"), links between users in Facebook, LinkedIn, etc. or semantic hierarchy information as in Wikipedia. Some researchers have already taken into account this information to enhance traditional text mining tasks and even develop new ones, such as the extraction of interest-based communities [PPZ+12] or the distinction between actual news and rumours from microblogging messages [MPC10].

Being aware of these limitations and opportunities, researchers have adapted traditional text mining techniques to this new scenario. This is the case of Hong and Davison [HD10], who analyse the performance of standard topic models (LDA) on social media data. Despite the difficulties in obtaining a Facebook dataset with user generated content, some researchers have taken advantage of this online social network to mine users' interests. This is the case of Jin et al. [JWL+11], who propose a system to infer users' interests from *Like* activities in Facebook. Since using the *Like* function is an indubitable sign of interest, the analysis of these activities can provide a direct and simple knowledge base to mine users' interests, users' representativeness, users' influence and so on. Palsetia et al. [PPZ+12] propose extracting interest-based communities in Facebook considering, similarly to Jin [JWL+11], direct signs of common interests (contributing to a specific Facebook wall or to a Twitter Profile), constructing, in this way, a global network from the dataset. Zhao et al. [ZYL+12] opt for a clus-

tering approach for the classification of users' interactions according to a set of Activity Topics deployed over a Facebook dataset. Although this work focuses on tie strength estimation, interaction documents are processed with different NLP-techniques. However, Zhao's proposal [ZYL+12] requires an a priori consideration of the fields of activity, since clustering is carried out by LDA.

Facebook is not the only social site used to extract users' interests, but other media as Twitter, have also served as sources of data to mine interests. In 2009, Banerjee et al. [BCD+09] proposed inferring interests from tweets using unstructured text mining techniques. Specifically, and according to their content, authors classified tweets into (i) ephemeral (the interest in an activity changes over time), (ii) descriptive (the interest can be described using one or more indicative keywords), and (iii) localised (the interest is associated with location information). Their strategy is based on extracting a list of keywords from a huge dataset of tweets and applying statistical techniques to discover associations between these keywords. Later, O'Connor, Krieger and Ahn [OKA10] presented *TweetMotif*, an exploratory search application for Twitter that groups messages by frequent significant terms. Their topic extraction mechanism is based on NLP techniques for syntactic filtering, scoring and filtering of topic phrase candidates, merging similar topics and grouping near-duplicate messages. Similarly, and with the aim of providing a more organised view of the user's feed, the Twitter client *Eddi* proposed by Bernstein et al. in 2012 [BSH+10] groups tweets in the user's feed into topics mentioned explicitly or implicitly. *TweeTopic*, the topic assignment algorithm of *Eddi*, proceeds in three steps: text transformation, search engine querying and result mining. The central intuition behind *TweeTopic* is that the brevity of tweets forces us to use an external knowledge base to expand our knowledge about the tweet. However, *Eddi* puts the emphasis on assigning a topic to a tweet, instead of extracting the interests of a specific user from his tweets. LDA [BNJ03] has also been applied in the literature to the problem of classifying user profiles in Twitter. With *TweetLDA* [QAC12], Quercia et al. propose a supervised topic classification method (based on LDA) for the task of document classification in Twitter – given a Twitter profile and a set of possible topics, determine which topics best fit the profile's tweets. Similarly, *Topick* [DOMA12] automatically detects Twitter users' interests in a set of predefined high-level topics by using LDA and a pre-computed topic model. Although LDA

is a well-known algorithm to extract topics from text, the fact that (i) topics have to be previously fixed, (ii) collections of interactions between users do not form a structured text and (iii) the usually short length of the user generated content, make it not the most suitable option for mining users' interests from social media data. Conversely, the aim of clustering is to arrange a collection of data into a small number of groups of similar elements without providing a set of topics or classifiers in advance. Clustering has also been applied to the problem of topic extraction in Twitter. In [RKT11], Rangrej et al. compare the performance of various clustering techniques on short text data collected from Twitter, concluding that Affinity Propagation performs better than K-means (similar to Partition Around Medoids), but without providing any supervised measure that takes into account human judgement. Kang et al. [KLP10] have also applied Affinity Propagation to tweets analysis, but once again the aim is slightly different: extracting clusters of tweets and not of tags representing interests. Their relatedness measure is based on syntactic matchmaking and frequencies of syntactically identical words, which is a serious limitation in social media where syntactic rules are usually relaxed.

2.1.3. Applications

Being able to extract users' interests from user generated content in social media has several applications, especially in the personalisation domain. One of the foremost applications in marketing is allowing the appropriate identification of "target marketing", that is, the group of customers towards which a business should aim its marketing efforts and ultimately its merchandise to maximise its benefits. This is the case of the social publicity tool *Facebook Ads* [Facb] that Facebook uses to disseminate advertisements (ads) among its users. However, the fact that (i) the recommendation algorithm, (ii) the users' recommendation profiles and (iii) the set of all available ads are not publicly available prevents it from serving as a reference to compare with any other social publicity strategy that considers Facebook users' data. Still, some recent work in the literature has inspected the use of the Facebook API for item recommendation [SRF13, DPH12, AHH+13, GKBM11, AGHT11]. The main focus of this work is to solve well-known issues in recommender systems, such as the cold-start and sparsity

problems, by enriching users' profiles with the aggregation of data gathered from different social networks. However, they mostly consider information consciously provided by users through different forms, tags, categories, etc, without taking into account the user generated content in their posts (comments, messages, etc.).

Apart from recommending products, other studies have focused on recommending contacts to link with in the network, such as that of Pennacchiotti and Gurumurthy [PG11] in which they propose a system to recommend to a user new friends that share similar interests. To this aim, they extract users' interests using LDA. Strategies to extract topics from tweets have also been used to identify influential users. In [WLJH10], Weng et al. present *TwitterRank*, an algorithm to identify influential users in Twitter by taking into account both the topical similarity between users and the link structure. In order to extract topics from users' tweets, they use LDA. Topic detection is also one of the techniques used by Cataldi et al. in [CDCS10] to retrieve emergent topics in Twitter. Their proposal is based on extracting terms from tweets and considering "emerging" terms those that often occur in a given time interval and were relatively rare in the past. They also take into account (i) the authority of the user (obtained by analysing the social relationships in the network) and (ii) a navigable topic graph that connects the emerging terms with other semantically related keywords, allowing the detection of the emerging topics.

2.2. Tie strength

2.2.1. Sociological perspective

The concept of *tie strength* was introduced in 1973 by the anthropologist Mark Granovetter in his iconic paper "The Strength of Weak Ties" [Gra73]. In it, he defines tie strength as a function of duration, emotional intensity, intimacy and exchange of services. He distinguished two kinds of ties in social networks: strong and weak ties. We keep strong ties with people that we really trust, often people like us (homophily), whereas weak ties relate us with simple acquaintances. Granovetter developed his tie strength framework in the context of job-hunting,

highlighting the importance of weak ties for individuals' integration into communities since they act as bridges between otherwise unrelated social clusters. Later, Granovetter revisited his theory [Gra83] with a round-up of studies that adopted tie strength, including the study in which Friedkin [Fri80] systematically demonstrates Granovetter's theory. Granovetter also shows that the notion of strength of a tie depends on the context and the nature of the tie. White holds a similar position in [Whi08], postulating that a social network is composed of different subnetworks depending on domains (*Netdoms*). Dunbar, in his well-known social brain hypothesis [Dun98], goes deeper by stating that the cognitive constraints of the human brain limit the number of social relationships maintained by a person at different levels of emotional closeness, this number being around 150 (Dunbar's number). Specifically, Dunbar postulated the existence of four "circles of acquaintanceship" (Dunbar's circles), which are, from the closest to the weakest: *support clique*, *sympathy group*, *affinity group* and *active network*. Later, he, in collaboration with Sutcliffe and others [SDBA12], found that the distribution of people in these circles is not constant, but the ratio between the sizes of two successive circles is almost constant, and very close to 3.

In sociology, apart from theoretically defining tie strength, most research has focused on substantive applications, such as the efficacy of weak ties in job search efforts ([Gra95, LD86]) or in the integration of scientific communities ([Fri80]). These studies were mainly conducted through surveys of human participants, providing only a limited and very static view. Although many studies have dealt with assessing tie strength and its implications in sociology, this does not aim to be a review of all the tie strength studies in sociology, but just a motivation that serves as a basis to assess tie strength in social media.

2.2.2. Tie strength in social media

Although the importance of tie strength was recognised with the publication of [Gra73] more than four decades ago, the emergence of social media and its widespread use have made this concept more relevant and important than ever, rolling out their study and importance to other disciplines beyond Sociology such as Computer Science. Even before online social networks were mainstream, Mut-

ton [Mut04] had already proposed a method for inferring a social network by monitoring an IRC channel in which, to obtain the network, an IRC bot observes the messages exchanged between users in the channel and, from this information, infers the social network in which they are involved. Another notable example is the work of Tyler et al. [TWH05], who proposed a method for identifying communities using e-mail data.

In the case of social media sites, most online social networks provide users' social structures composed of links between users and their contacts on these networks. These links or relationships are usually considered equal, but this is not the case in reality, nor in an online environment: a person can only maintain a limited number of social relationships at different levels of closeness [Dun98]. Therefore, several computer scientists have focused on studying the interaction network (networks made of ties between users who often interact through social networks) and, specifically, on highlighting its enormous differences from the social network provided by the site. For example, Kahanda and Neville [KN09] studied how to infer the nature and strength of relationships between Facebook's members using attribute-based features (gender, relationship status, ...), topological features (connectivity of the users in the friendship graph), transactional features (wall postings, picture posting and groups) and network-transactional features (Wall posting in another user's wall, ...) to obtain users' "top-friends". Using an application that allowed users to mark their "top-friends", they concluded that the most important features to predict tie strength were the network-transactional features, followed by the transactional ones. In a similar study, Gilbert and Karahalios [GK09] present a predictive model that maps social media data to tie strength. The model builds on a dataset of social media ties on Facebook in which, apart from considering interactions between users, they consider factors like age, political ideals or distance between hometowns. Xian et al. [XNR10] present a latent variable model for predicting tie strength based on profile similarity and interaction activity of users in the site, validating this model in Facebook and LinkedIn. Other studies, such as [WBS+09, VMCG09, BBK+11], also focus on mining users' activity on Facebook to calculate tie strength, taking into account different interaction signs. Wilson et al. [WBS+09] consider, for each user, the social graph, wall posts and photo comments as evidence of interaction, whereas Viswanath et al. [VMCG09] only take into account wall posts to

study how the varying patterns of interaction over time affect the overall structure of the interaction network. Finally, Backstrom et al. [BBK+11] study how Facebook users allocate attention across friends, taking into account, apart from messages, comments and wall posts, information about how many times one user views profile pages or photos posted by other users. Both Wilson [WBS+09] and Viswanath [VMCG09] use Facebook data obtained using crawlers, whereas Backstrom [BBK+11] retrieves data directly from Facebook, since information about users' passive interactions such as browsing updates, photos or profiles of their friends, is not publicly available. Although most of the previous studies have been focused on traditional social networks as Facebook, there have been some researchers who studied tie strength in other platforms, such as on the microblogging server Twitter. This is the case with Huberman et al. [HRW09], who infer which Twitter followees (followers) are truly related to the user by taking into account directed tweets (mentions in tweets). Grabowicz et al. [GRM+12], meanwhile, validate Granovetter's hypothesis [Gra73] in online social networks, so that links with retweets should be more likely to appear as bridges between different groups, whereas links with mentions should connect users in the same groups.

2.2.3. Effects and applications of tie strength

Apart from defining and assessing tie strength both from surveys or online social media sites, research on tie strength has also focused on analysing the effects of tie strength on users' behaviour as well as how it may be used in technological applications. Below, we review some examples of studies that have touched on these two areas.

In 2007, Onnela et al. [OSH+07] carried out one of the first large-scale studies analysing the effects of tie strength on information diffusion within a network. They study the social network obtained from a large dataset of mobile phone calls, where a phone call represents a link between two users and the aggregated duration of their calls determines their tie strength. They also simulate the spread of information through the network, finding that both weak and strong ties are ineffective when it comes to information transfer, the former because the small

amount of on-air time offers little chance of information transfer and the latter because they are mostly confined within communities, with little access to new information. Later, Bakshy et al. [BRMA12] examined the role of strong and weak ties in information propagation within an online social network. Using an experimental approach on Facebook, they found that, although stronger ties are individually more influential, weak ties are responsible for the propagation of novel information. This supports Granovetter's theory [Gra73] in the online world, since these findings highlight the importance of weak ties in the dissemination of information online. In another related study, Zhao et al. [ZWF+12] investigated the impact of tie strength on information propagation in online social networks. They observed that, compared with weak ties, strong ties are more favourable for information diffusion in OSNs (Online Social Networks), but they alone are not adequate for widening the spread of information. They went a step further by distinguishing between "positive" and "negative" weak ties, the former being the ones with nodes that are centres of different clusters or local communities, and the latter the ones that contain low-degree nodes with very small overlaps of friends. They found that although the negative weak ties tend to hinder information from being further diffused, the positive weak ties have an important bridge effect that can facilitate information propagation across various isolated communities. More recently, Karsai et al. [KPV14] analysed rumour-spreading processes in a mobile call dataset, and the effect of strong and weak ties in their propagation. They found that strong ties have an important role in the early cessation of rumour diffusion by favouring interactions among users already aware of the gossip. That is, strong ties constrain the diffusion of information by confining the spreading process, having a negative role in the spreading of information across networks. In order to carry out this study, they took into account the microscopic dynamic evolution of the network inferred from the mobile call dataset.

Contrary to those studies that focused on the whole network to analyse the diffusion of information, other studies have focused on the ego network of the user in order to see how users get the most useful information and novel information: from weak or strong ties. Some examples are the work of Panovich et al. [PMK12], who studied the role of tie strength in question answers within online social networks and the relation between tie strength and the quality of the answer. They conducted a user study using Facebook as the social network under

consideration, finding that stronger ties (close friends) provide better information and share less information that the participant already knows than weak ties. In the same vein, Burke and Kraut [BK13] found that after losing a job, strong ties were more useful to find new employment within 90 days. Both Panovich and Kraut findings are somewhat unexpected, since they contradict Granovetter's theory [Gra73]. On the other hand, in a similar study to that by Panovich, Gray et al. [GEVL13] concluded that useful responses are more likely to be received from weak ties on Facebook.

Tie strength also has impact in marketing. In a 2006 study, Hill et al. [HPV06] proved that links between customers can directly affect product/service adoption. Using data about the adoption of a new telecommunication service, they showed that those customers linked to a prior customer adopted the service at a rate 3-5 times greater than baseline groups selected by the product's marketing team. Later, Wen et al. [WTC09] investigated the impact of the tie strength between the consumer and the endorser, the consumer's perception of the endorser, and the type of product on effective product endorsement on social network sites. By means of a 201-user study, they found that for hedonic products (designer clothes, sports cars, luxury watches, etc.) strong-tie endorsers are more effective than weak-tie endorsers, regardless of their expertise on the product; for utilitarian products (microwaves, minivans, personal computers, etc.) high-expertise endorsers result in higher consumers' purchase intention, regardless of their tie strength with consumers.

There have also been studies focused on socially-enhancing technological applications using tie strength, especially in the area of recommendation. This is the case for [SC11], where Sharma and Cosley present *PopCore*, a Facebook application to conduct experiments in network-centric recommendations. They propose six different recommendation algorithms to explore the effects of popularity, personalisation, similarity and tie strength on the recommendation of movies, television and books. By means of a 50-user study, they found that the best recommendations are provided when taking into account popularity (especially when the items are popular among users' friends), followed by the algorithms that consider tie strength and similarity. Non-personalised algorithms are the ones that provided the least accurate recommendations. Also involving

Facebook, Chen et al. [CF10] propose enhancing the collaborative filtering approach in recommender systems by taking advantage of trust between users in online social networks. In order to assess this trust they consider, besides other components, the strength of their relationship obtained mainly by mining users' interactions on Facebook. Later, and concerning Twitter rather than Facebook, Chen et al. [CNC11] analysed different algorithms to recommend conversations to Twitter users using, among other components such as thread length or topic of the tweet, the strength of the tie between users. By conducting an online user study, they concluded that the five different recommendation algorithms they proposed suggested more interesting conversations than a random baseline, and that the tie strength-based algorithms performed better for people who used Twitter for social purposes than for only being informed. Apart from improving traditional recommendation algorithms, Gartrell et al. [GXL+10] demonstrate the benefits of using tie strength to improve group recommendations. They propose a group recommendation algorithm that considers, apart from the interests of the group members, information about the strength of the relationship between the members within the group. Considering five different levels of closeness (strength), they conducted a user study in which participants had to assess the strength of their relationship with others, rate movies according to their preferences and discuss the movies with others to get group ratings for the movies. Although Gartrell et al. accurately predicted some of the group decisions, their algorithm failed to predict the group consensus in others.

Tie strength has also been successfully used in other socially-enhanced services. As early as in 2001, Cortes et al. [CPV01] applied a dynamic network model based on interactions between users (phone numbers) to detect different types of fraudulent behaviour in a telecommunication network, such as subscription fraud (when an account is set up by a user who has no intention of paying any bill) or when a fraudster has assumed a new identity. In their 2009 extensive analysis of users' interactions on Facebook, Wilson et al. [WBS+09] showed that the interaction graph (network made of links between users that really interact on Facebook) can successfully be applied to two different socially-enhanced applications: collaborative spam mitigation [GKF+06] and defence against Sybil attack [YKGF06]. Specifically, Wilson demonstrated that the use of interaction graphs extracted from Facebook improves the performance of "RE", the

white-listing system for email based on social links that allows emails between friends and friends of friends to bypass standard spam filters proposed by Garris [GKF+06]. He also proved the benefits of considering the interaction network to detect Sybil identities in an online community in order to protect distributed applications (Sybilguard [YKGF06]).

2.3. Open problems to address

In this chapter we have reviewed existing approaches to discovering interests and social ties from user generated content in social media sites and described how these data have been used to personalise or socially-enhance services. As previously indicated, the fact that applications to personalise consider data retrieved from only one social media site and that often this site is also the medium to deliver the application are some of the main shortcomings of previous research in using social media data to personalise applications. In order to make up for these deficiencies, in this dissertation we define a user-centred model for personalising applications based on mining data from different social web sites to identify the subjects – topics – of interests of users (social contexts) and the strength of the relationships between them and all the users with whom they interact (tie strength with their social contacts or contacts that form their social spheres).

Specifically, we build on previous work by presenting a methodology to extract interests from users-generated content in social media sites, together with the results of its evaluation. Unlike the work thus far described, our methodology can accurately extract, represent and manage interests extracted from any social site no matter the length of users' textual publications; and deliver these interests to any service that users have granted access to. In addition to the social publicity application that we personalise using the interests extracted and that we also describe, we discuss how to personalise other applications (both existing or to be created). Also here, we discuss how a large scale prototype of an intermediary service that builds and delivers social spheres might be designed and implemented and the technological issues that still need to be addressed before such a service could be adopted as a technological tool for personalisation.

The other key element needed to build the social spheres is the tie strength measure. While previous work provides models to develop an interaction network, our proposal focuses on the individual, providing a user-centred measure of the relationship between two social media users. We build on previous works by presenting a measure of the strength of the tie between two individuals that interact through social media tools from the perspective of one of them, the user for whom the sphere is provided. Unlike some previous proposals, our measure only considers interaction data retrieved through public APIs with users' permission. It also takes into account the time at which interactions occur, since recent interactions may have a greater impact than older ones, and the people involved in them, to assess the relevance of the person in the interaction. As in the interests extraction proposal, we also discuss how to personalise applications using the tie strength between users obtained with our measure in particular, and the *contextualised social spheres* in general. That is, spheres composed of those users with whom the target user frequently talks about the topic – context – of the sphere, together with the strength of their ties taking into account only those interactions in the scope of the topic.

3

A tag clustering-based approach to extract social interests

Social Web technologies have triggered the evolution of Internet users from being mere observers of the continuous generation of information to become active and prolific producers of content. This user generated content usually contains evidence of individuals' interests, feelings and likes. Although traditional text mining techniques have successfully handled the extraction of topics from documents (Chapter 2), the fact that data generated on social media sites are considered vast, noisy, distributed, unstructured and dynamic [GL12], presents very different characteristics from the usual attribute-value data used in classic data mining. Therefore, new techniques have to be applied to reduce the harmful impact of noise or to deal with managing unstructured data from different data sources in order to successfully extract interests – topics – from individuals' publications in social media sites.

In this chapter, we address the problem of mining users' interests from user generated content on social media sites. Our mining process uses different data mining and natural language processing techniques to obtain clouds of tags (bags of words) representative of users' interests. This solution can be developed without any a priori knowledge about the number and category of interests, nor a priori knowledge about the users for whom extraction is applied. To show how our solution works, we describe a deployment scenario for social publicity where knowing users' contexts – interests – allows advertisements to target potential customers. Using *Groupon* [BMPZ11], a deal-of-the-day website, as a source of ads and Facebook as social knowledge source, we provide users with personalised ads recommendations according to their contexts. Also, by their publication on Facebook walls, ads would be spread by word-of-mouth power throughout users' contacts. An important advantage of our model is that it works independently of the ads source (*Groupon* or others), since we do not consider any prefixed categorisation (fashion, sports, traveling, etc.). Instead, we focus on finding the best ad by NLP-analysing the textual descriptions of the ads source. We also describe how we validated our proposal by means of a user study using the previous deployment scenario. Results of this study revealed that comparing resulting clusters against human judgement does not work fine for our purpose since users are not as reliable as expected when they are asked about their interests. For this reason, we conducted a second experiment that, instead of gathering feedback by explicit contribution of users, uses the implicit classifying method used by Twitter users to organise their tweets: *hashtags*. This experiment also allowed us to evaluate the performance of three different clustering algorithms.

In relation with our thesis, our findings suggest that the proposed methodology can accurately extract individuals' interests from user generated content in social media, and the fact that these interests are represented by bags of words might be used to personalise almost every application.

3.1. A model for extracting users' interests

Interest similarity-based community detection, i.e. finding groups of people with shared interests, has received a lot of attention in the research community

(see [PRS11] for a comparative study). However, our work has a slightly different focus: our intent is not to construct a supra-network or meta-network from a social network as in *Community Detection*, but to infer (i) the topics talked about by a social media user – social contexts – and (ii) the friends with whom he talks about each specific topic – friends belonging to each social context – in a local way (*Local Community Detection*). In short, we do not aim to identify which users deal with a topic – find communities of users with common interests–, but which topics are treated by a given user.

As indicated in the previous chapter, a good algorithm to extract topics from text is LDA. However, the fact that (i) topics have to be previously fixed (not useful in this case) and (ii) collections of interactions between users do not form a structured text, makes it not the most suitable option for our task. Although this scenario is different from traditional collaborative tagging systems [Bur02, AT05, GH06], we take advantage of the techniques used in these systems, and specifically of the use of keywords or tags to define users' contexts. With tools like *Stanford Core NLP* [MSB+14] and *Freeling* [CCPP04] that provide different NLP mechanisms, such as Part-Of-Speech (POS) Tagging, getting a word citation form (lemmatisation), Named Entity Recognition (NER), etc., it is possible to extract meaningful information from texts in natural language.

The quality of the extracted information clearly improves if the semantics of the words and the relations between them are considered. Many tagging systems obtain this semantic relatedness by checking how many times two tags appear together tagging the same item [FDS14]. However, the absence of data about all the users in social media services, that is, a global folksonomy, makes it unfeasible in this case. So, we have explored other possibilities to measure the semantic similarity between tags, whose main difference is the source of background knowledge by which they are supported. Relatedness measures based on the handcrafted lexical database of English *WordNet* [PPM04], on the Web (WWW) [CV07] or on Wikipedia [SP06, GM07, WM08] are only some of the measures proposed. Since online users usually talk about trending topics, which include people who have become famous recently, new TV programs, new products, etc., measures of semantic relatedness between terms should be based on frequently updated sources of background knowledge such as Wikipedia or the Web instead of on

traditional lexical databases as *WordNet*.

The last enabler to our proposal is clustering. This technique of exploratory data analysis aims to identify patterns in data by the discovery of groups of strongly related data points. That is, clustering is about splitting a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters. A huge number of clustering algorithms have been proposed to date, which can be classified according to different criteria (see [JMF99] for a review). Most of the existing clustering algorithms can be labeled either hierarchical or partitional. The difference lies in that a partitional clustering algorithm obtains a single partition of the data instead of the clustering structure, known as a *dendrogram* (see [JMF99] for an example), which represents the nested grouping of patterns and similarity levels at which groupings change and which is the result of a hierarchical clustering.

In this section we present our methodology for extracting topics from user generated content in social media sites, i.e. users' interests manifested in these media. This methodology also identifies the friends of the user with whom he shares these interests. We coined the term "*social context*" of a user to refer to the set of interests and friends that share these interests with the user and, as in real life, users have different interests and even different numbers of them. Our extraction methodology is based on using textual descriptions linked to users' interactions (private message content, photo descriptions, etc.), gathered from social media sites (with users' permission). We then apply several Natural Language Processing (NLP) techniques to get the most representative words or tags from these data to build the personomy (the user's folksonomy). The user's social contexts emerge after applying a clustering algorithm over this personomy and classifying the user's contacts in these clusters. We detail this methodology in Section 3.2.

3.2. Methodology

The main contribution in this chapter is a model that, taking advantage of users' textual publications in social media sites, infers the social contexts in which

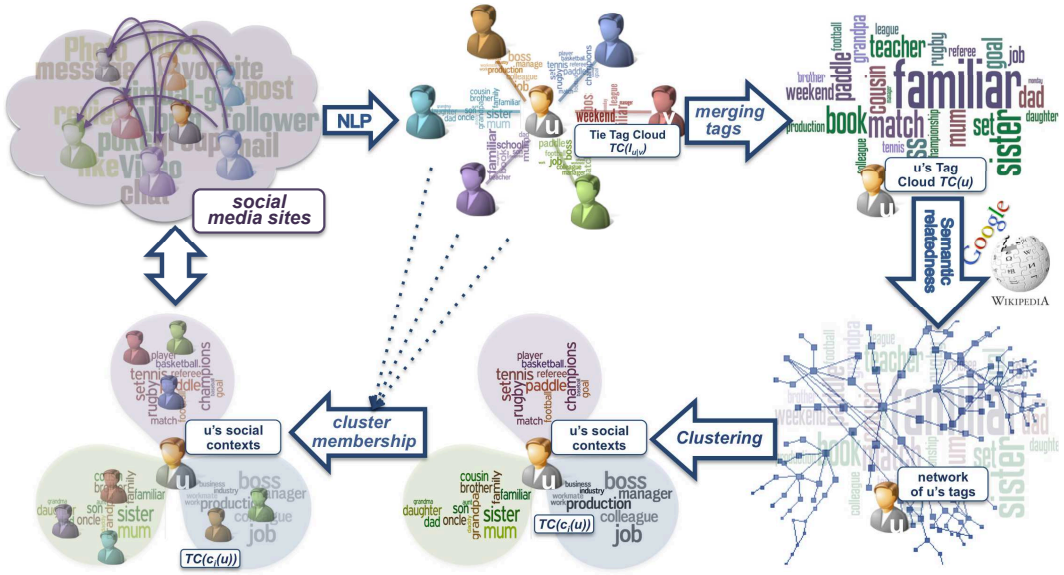


Figure 3.1: Inferring a user's social contexts

users are involved and which of their contacts belong to them. Figure 3.1 shows the different steps that form the model. After applying different natural language techniques (POS tagging, stop words removal, etc.) over data extracted from social media sites' APIs (*Application Programming Interfaces*), we obtain the user's personomy, that is, the tags that represent his online life and the semantic relationships between these tags – his network of tags –. The next step deals with applying a clustering algorithm to obtain groups of strongly related tags in the personomy, which will give rise to social contexts. Finally, the membership of each contact in a specific context is determined by the similarity between the tags of the context and the tags linked to the social interactions between the user and the contact.

Although in the figure each publication is linked to only one tie between the user u and one of his contacts, other situations are possible. Publications can be associated with more than one tie (with more than one contact) or even with all the ties – or none of them - when the user addresses the content to his whole audience, without expressly indicating any addressee. In this case, it would be more precise to talk about the user's *individual* interests – or contexts – rather than social contexts.

The following sections describe (i) how to obtain a user's personomy, (ii) how

to infer social contexts from this personomy and (iii) how to classify contacts into contexts.

3.2.1. User's Personomy

A user's personomy is built from the tag cloud representative of the user's interactions and the semantic relations between these tags.

- **User's Tag Cloud**: We use the textual descriptions of the interactions between a user u and each one of his contacts v (wall posts, private messages content, photos descriptions, etc.) retrieved from social media sites' APIs to obtain the tag cloud which characterises u's link with each contact v, $TC(l_{u|v})$. Obviously, these descriptions are not provided in form of tags. To extract the relevant words (tags) we only consider lexical units that refer to fixed entities with meaning. To this aim, we use tools like *Stanford CoreNLP* [MSB+14] or *Freeling* [CCPP04] that address all the basic levels of NLP[1]. We start filtering the text, using POS tagging (which identifies each word part-of-speech category – noun, verb, etc.) and lemmatisation (which identifies each word lemma or citation form). After only keeping nouns and verbs in their citation form, we remove the *stop words*, i.e. extremely common words such as *be*, *thing*, etc. The resulting words form the set of tags of the social interactions between u and v, $T(l_{u|v})$, whose tag cloud is denoted by $TC(l_{u|v}) = \{t, w(t, l_{u|v})\}$; $w(t, l_{u|v})$ being the importance of the tag t in the set (percentage of occurrence):

$$w(t, l_{u|v}) = \frac{m(t, T(l_{u|v}))}{\#T(l_{u|v})} \tag{3.1}$$

where $\#T(l_{u|v})$ is the number of tags in the relationship between u and v and $m(t, T(l_{u|v}))$ is the multiplicity of the tag, i.e. the number of times that this tag was used in the interactions between u and v.

A user u is modelled by $T(u)$, the set of tags resulting from merging all the

[1]Note that *Stanford CoreNLP* only analyses text in English. Other tools, like *Freeling*, also provide support for other languages, including Spanish.

tags that characterise the user's links with every contact $v \in Contacts(u)$. That is, the set of tags representative of the user's online life:

$$T(u) = \bigcup_{v \in Contacts(u)} T(l_{u|v}) \tag{3.2}$$

Analogously, we define the tag cloud of a user u as follows: $TC(u) = \{t, w(t,u)\}$; $w(t,u)$ denotes the importance of the tag t in the set (percentage of occurrence):

$$w(t,u) = \frac{m(t, T(u))}{\#T(u)} \tag{3.3}$$

where $\#T(u)$ is the number of tags of u's social publications.

- **Semantic Relationship Between Tags**: In order to discover groups of strongly related tags (clusters), we need to assess the semantic relatedness between the user's tags. To this aim, we take advantage of external sources of background knowledge such as the whole Web or Wikipedia and semantic relatedness measures proposed in the literature [CV07, SP06, GM07, WM08]. Some of them will be described later in Section 3.2.3.

3.2.2. Social Contexts

The social contexts of a user represent the different topics that he talks about in his interactions with others. People usually talk about the same topics with the same subset of contacts. Therefore, our goal is (i) to identify groups of strongly related tags, i.e. users' social contexts and (ii) to find out to which contexts their contacts belong.

- **Social Contexts Inference**: In order to identify these social contexts, we apply clustering over the user's personomy. To this aim, we take advantage of one of the huge variety of clustering techniques proposed in the literature [JMF99]. Analogously to the user's tag cloud, each context, c_i, is characterised by a set of tags (the tags in the corresponding cluster). The

context tag cloud is: $TC(c_i) = \{t, w(t, c_i)\}$; $w(t, c_i)$ denotes the importance of the tag t in the set:

$$w(t, c_i) = \frac{m(t, T(u))}{\#T(c_i)} \tag{3.4}$$

where $T(c_i)$ is the set of tags linked to the u's i-context, $\#T(c_i)$ is the number of tags in this set and $m(t, T(u))$ is the multiplicity of the tag, i.e. the number of times this tag was used in the interactions between u and his contacts. The resulting clusters have different numbers of tags and not all of them are representative enough to be a context. To model users' social contexts, we only keep those M clusters where the weighted sum of the tags in the cluster is higher than a threshold ($Th_{context}$).

- **Contacts Membership to Social Contexts**: Once the set of u's contexts have been obtained and characterised: $TC(c_i) \mid i \in [1, M]$, our objective is to identify u's contacts involved in each one of these contexts, i.e. those who talk to u about *football*, *family*, etc. To this aim, we fix a threshold ($Th_{membership}$) for membership in a context and compare the tag clouds $TC(l_{u|v})$ and $TC(c_i)$ to obtain the set of contexts $c_{u|v}$ of u in which a contact v is involved:

$$c_{u|v} = \{c_i \mid Similarity(c_i, l_{u|v}) > Th_{membership}\} \tag{3.5}$$

where $Similarity(c_i, l_{u|v})$ is the cosine similarity [FDS14] between $T(l_{u|v})$ and $TC(c_i)$. When computing cosine similarity, any tag cloud can be thought of as a vector whose components are the weights of all the tags and, consequently, it is just calculating the angle between both vectors.

3.2.3. Semantic relatedness and Clustering

3.2.3.1. Semantic relatedness measures

Different approaches have been proposed to assess the semantic relatedness between two terms, whose main difference is the source of background knowledge by which they are supported. As mentioned, given that online users usually talk about trending topics, measures of semantic relatedness between terms should

be based on frequently updated sources of background knowledge instead of on traditional lexical databases. For this reason, we selected a measure based on the Web –*Normalized Google Distance (NGD)* [CV07]– and another based on the hyperlink structure of Wikipedia –*Wikipedia Link-based Measure (WLM)* [WM08]– to use in our experiments. We briefly describe these measures below.

- ***Normalized Google Distance (NGD)*** [CV07]: Cilibrasi and Vitányi proposed the Google Similarity Distance, a method to automatically extract the similarity of words and phrases from the Web using Google page counts. NGD computes the similarity between two words (or phrases) from the number of hits returned by the Google search engine. The result is a numeric value, $NGD(x, y)$, that represents the degree of similarity distance between the words x and y:

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \tag{3.6}$$

  where $f(x)$, $f(y)$ and $f(x, y)$ denote the number of pages containing x, the number of pages containing y and the number of pages containing both x and y, as returned by Google. N is a normalisation factor which has to be higher than $f(z)$ $\forall z$ and whose value we chose equal to the highest number of results on Google ($N = 25,270,000,000$ at the time of the experiments).

- ***Wikipedia Link-based Measure (WLM)*** [WM08] is a semantic relatedness measure based on the hyperlink structure of Wikipedia. After identifying the Wikipedia articles that discuss the words of interests, the relatedness between the given articles is computed by means of two different measures: one based on the links extending out of each article and the other on the links made to them. The first measure is defined by the angle between the vectors of the links found within the two articles of interest (a, b), where the weight w of the link $s \to t$ ($s \in \{a, b\}$) is:

$$w(s \to t) = log(\frac{|W|}{|T|}) \mid \text{if } s \in t,\ 0 \text{ otherwise} \tag{3.7}$$

  where T is the set of all articles that link to t and W the set of articles in Wikipedia.

The latter is based on the *Normalized Google Distance* [CV07]:

$$sr(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \tag{3.8}$$

where A and B are the sets of articles that link to the articles of interest a and b respectively, and W is the entire Wikipedia. See [WM08] for a complete description.

3.2.3.2. Clustering algorithms

Clustering is one of the dominant techniques of exploratory data analysis whose aim is to identify patterns in data by the discovery of groups of strongly related data points. From all the clustering algorithms proposed to date we selected three to use in the experiments proposed in this dissertation: (i) the popular *Partition Around Medoids (PAM)* algorithm (partitional) for which the number of resulting clusters must be known in advance, (ii) *Affinity Propagation* [FD07] (partitional), that does not previously need to know the number of clusters, but the clusters and their exemplars emerge from a message-passing procedure and (iii) the hierarchical clustering algorithm known as *The Unweighted Pair Group Method with Arithmetic Mean (UPGMA* [JD88]). Next, we briefly describe each of them.

- ***Partitioning Around Medoids (PAM)*** [KR09] is a partitional clustering algorithm based on finding representative objects (medoids) in clusters, i.e. the most centrally located object within the cluster. The main steps of PAM are as follows:

  1. Randomly select k of the n data points as the medoids.
  2. Associate each data point with the closest medoid.
  3. For each medoid m and each data point o associated with m swap m and o and compute the total cost of the configuration (that is, the average dissimilarity of o to all the data points associated with m). Select the medoid o with the lowest cost of the configuration.

4. Repeat alternating steps 2 and 3 until there is no change in the assignments.

The PAM algorithm requires three parameters: number of clusters k, cluster initialisation and distance metric.

- ***Affinity Propagation (AP)*** [FD07] takes as input measures of similarity between pairs of data points and, by the exchange of real-valued messages between them, a set of *exemplars* (centres of the clusters) and their clusters gradually emerge. Rather than requiring that the number of clusters be pre-specified, Affinity Propagation takes as input a real number $s(k,k)$ – preference – for each data point k so that data points with larger values of $s(k,k)$ are more likely to be chosen as exemplars. The number of resulting clusters is influenced by the values of the input preferences, but also emerges from the message-passing procedure.

  Initially, all points are considered as potential exemplars, though each point can be manually assigned a *preference* that it should be chosen as an exemplar. For each point i and each candidate exemplar k, AP computes the *responsibility* $r(i,k)$, which indicates how well suited k is as an exemplar for i, and the *availability* $a(i,k)$ reflecting the evidence for how appropriate it would be for i to choose k as its exemplar:

$$r(i,k) \leftarrow s(i,k) - \max_{k':k' \neq k} \{a(i,k') + s(i,k')\} \tag{3.9}$$

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i':i' \notin \{i,k\}} \max\{0, r(i',k)\}\} \tag{3.10}$$

  where $s(i,k)$ denotes the similarity between the two data points i and k. The above two equations are iterated until a good set of exemplars emerges. Each point i can then be assigned to the exemplar k which maximises the sum $a(i,k) + r(i,k)$, and if $i = k$, then i is an exemplar.

- ***Unweighted Pair Group Method with Arithmetic Mean (UPGMA)*** [JD88] is a hierarchical and agglomerative clustering algorithm that yields a dendrogram that can be cut at a chosen height to produce the desired number of clusters. The main steps of UPGMA are as follows:

1. Place each data point into its own singleton group.
2. Merge the two closest groups.
3. Update distances between the new cluster and each of the old clusters. Given a distance measure between points, UPGMA obtains the intergroup similarity between the clusters C and H as:

$$d(C,H) = \frac{1}{N_C N_H} \sum_{i \in C} \sum_{j \in H} d_{i,j} \tag{3.11}$$

   where N_C (N_H) is the size of the cluster C (H) and $d_{i,j}$ is the distance between the data points i and j.
4. Repeat 2 and 3 until all the data are merged into a single cluster.

As with any hierarchical clustering, UPGMA only requires a measure of similarity between groups of data points.

3.3. Experiment 1: deals recommendation on Facebook

Our second contribution in this chapter is a deployment scenario for social publicity where the knowledge of users' contexts allows advertisers to target potential customers. Apart from being an example of how to apply our methodology to real scenarios, an early stage implementation of this scenario allowed us to validate our methodology against human judgement. This application takes advantage of *Facebook* [faca], the online social networking site par excellence, and *Groupon* [gro], a deal-of-the-day website which offers discount coupons usable at several companies, to provide users with personalised advertisements. Facebook is a popular online social networking site launched in 2004 that provides users with the typical interpersonal communication features: posting on walls, exchange of private messages, uploading and sharing photos, etc, which usually have some associated textual content (photo title, photo comment, wall-post content, private message content, etc). Users may personalise the privacy of these activities to restrict access to profile information, mini-feed, wall posts, etc. (only to friends,

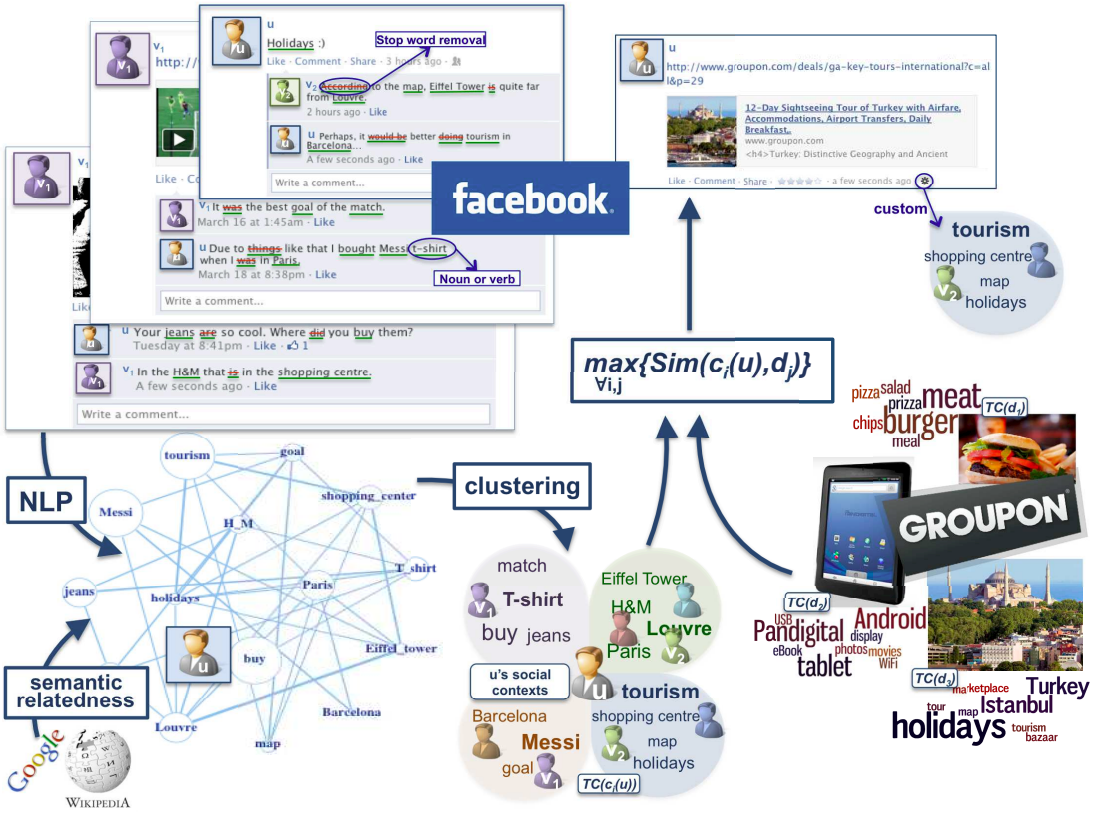


Figure 3.2: Personalised publicity on Facebook

friends-of-friends, lists of friends, no one and all). Facebook provides developers with an API through which third party applications can access users' data on Facebook on their behalf (with their permission) and even post content on their walls. *Groupon*, the deal-of-the-day website launched in 2008, also provides an API for developers to obtain data about the deals and facilities to integrate these deals with external websites.

Technically, this application, whose overview is depicted in Figure 3.2, is in charge of (i) obtaining the user's social contexts (c_i) and the contacts who belong to them following the methodology proposed in the previous section, and of (ii) daily retrieving, from *Groupon* API, a short description of the deals of the day, d_j. For each deal, its tag cloud ($TC(d_j)$) would be built in the same way as the tag cloud of the user's social personomy described in Section 3.2.1. These deals' tag clouds ($TC(d_j)$) would be compared with the user's contexts tag clouds ($TC(c_i)$) using $Similarity(c_i, d_j)$, to get the deal with the highest similarity with some of the user's contexts. A link to this deal would be finally posted on the

user's wall. To avoid disturbing the user's contacts with excessive publicity, the post would only be visible to those contacts that belong to the selected context and, therefore, are probably interested in the product.

Apart from an overview of the application, Figure 3.2 shows a toy example in which, after obtaining his social contexts and contacts, a user u receives, in the form of a wall-post visible only to the proper contacts, the ad that best matches his social life. In this toy example, the lemmatizer and the POS-tagging of a NLP tool have been used, after removing the stop words, to obtain the 15 tags that represent user u's social life (*tourism*, *Barcelona*, *T-shirt*, etc.). Then, clustering has been applied over u's personomy – previously calculated by a semantic relatedness measure and taking into account the weight of each tag – to obtain the four tag clusters that represent his social contexts and whose exemplars are *T-shirt*, *Louvre*, *Messi* and *tourism*. Note that the tag *match* belongs to the context whose exemplar is *T-shirt* instead of the context represented by *Messi*, which could be solved using, previously, some word disambiguation technique. Finally, the tag cloud of each context ($TC(c_i)$) is compared with the tag cloud of each deal ($TC(d_j)$) that, in this case, represents discounts in (i) a fast-food restaurant, (ii) the purchase of a tablet and (iii) a trip to Istanbul. The highest similarity is between the trip to Istanbul and the context whose exemplar is *tourism*. Note that v_2 is one of u's contacts in this context.

3.3.1. Problem definition

This section reports the experimental evaluation of our model to infer social contexts using the aforementioned deployment scenario. Our tests involved 30 (under)graduate students from the University of Vigo, their friends and relatives. We ended up with a diverse audience, with disparate demographic data and educational backgrounds, including nearly as many men as women with ages ranging from 21 to 48 years. We worked with a group of 10 different *deals* (obtained from the *Groupon API*) which included a wide range of discounts in products and services including a breathalyser, an online *Autocad* course, beer, an *ipad* (tablet PC), a hotel stay in Madrid, a pair of earphones, a football pool discount voucher, an online series subscription, pizzas and a desk chair. As interaction

data, we considered wall-posts exchanged, comments (including the description of the wall-post, photos, previous comments, etc.) and textual content associated with users' (contacts') items that contacts (users) *like*. In this situation, the evaluation consisted of (i) asking users to rank the deals according to their preferences, (ii) ranking the deals according to their similarity with users' contexts (interests) obtained using our methodology and, finally, (iii) checking the correlation between both rankings.

3.3.2. Particularising the model to this scenario

Our mechanism of interest extraction starts by obtaining users' publications from the Facebook API and the extraction of their relevant words by only considering lexical units that refer to fixed entities with meaning. To this aim, we used *Freeling* [CCPP04], a tool that addresses all the basic levels of NLP, since it works for Spanish, the mother tongue of the participants. Once users' tag clouds had been obtained, the semantic relatedness – distance in this case – between tags necessary for obtaining the users' personomy was computed using the *Normalized Google Distance (NGD)* proposed by Cilibrasi and Vitányi in [CV07]. Finally, the clustering algorithm *Affinity Propagation (AP)* was used to extract users' interests from this personomy. See Section 3.2.3 for a detailed description of these techniques.

3.3.3. Parameters estimation for training

In order to fix the different parameters of our methodology, we asked 11 students about the topics they talked about on Facebook and to select their 3 preferred deals. We repeated the steps of the process for getting their social contexts varying (i) the value of the parameter q that controls the exemplar preferences in *Affinity Propagation*, (ii) the threshold which determines if a cluster is representative enough to form one context ($Th_{context}$) and (iii) the threshold for contacts' membership ($Th_{membership}$).

The following shows the estimation of every parameter, where TP is the set of topics indicated by each user and CL the resulting clusters from applying Affinity

Propagation over his tag cloud $TC(u)$.

- In order to estimate q, we defined the following function

$$f(q) = \underset{i}{avg} \left\{ \frac{NGD(tp_i, cl_s)}{\underset{j}{avg} \{NGD(tp_i, cl_j)\}} \right\}, \tag{3.12}$$

that represents the average of the semantic distance between each topic pointed out by the user ($tp_i \in TP$) and its nearest cluster $cl_s \in CL$ with respect to the average distance between the topic and the rest of the user's clusters $cl_j \in CL - \{cl_s\}$. The optimal value of q is the one that minimises $f(q)$ and, in turn, maximises the number of different nearest clusters for different topics. In view of the results, and as aforementioned, the number of contexts (clusters) may be different for different users. Therefore, the parameter q, whose value is inversely proportional to the number of clusters, should depend on the user and specifically on the degree of semantic distance between all the tags in his tag cloud. As expected, the most suitable value of q is obtained when q is directly proportional to the average of all distances between the user's tags, the average value denoted by $\bar{X}$. We applied a linear regression to discover the relation between q and $\bar{X}$, obtaining that its relation is determined by $q = m\bar{X} + x_0$, being $m = -0.9129$ and $x_0 = 2.4431$ (see Figure 3.3).

- Having fixed the parameter q, not all the resulting clusters may be sufficiently relevant. We consider that a cluster is representative for a user u, and therefore forms one of his social contexts, when the sum of its tags with respect to the sum of all the user's tags

$$\sum_{t \in T(cl)} w(t,u) / \sum_{t \in T(u)} w(t,u) \tag{3.13}$$

is higher than $Th_{context}$. Figure 3.4 shows the variation of the recall and precision with variation of this threshold. In view of the results, we opted to use the value 0.03 as $Th_{context}$. Note that this value should depend on the application that uses the social contexts.

- Finally, we run the method to fix $Th_{membership}$ that determines the mem-

bership of a contact in a cluster of u (see Equation 3.5). This value should also depend on the application. In order to limit deal propagation, so that excessive social publicity will not be perceived as disturbing by users' contacts, we limited the number of contacts who would see the ad to three per context.

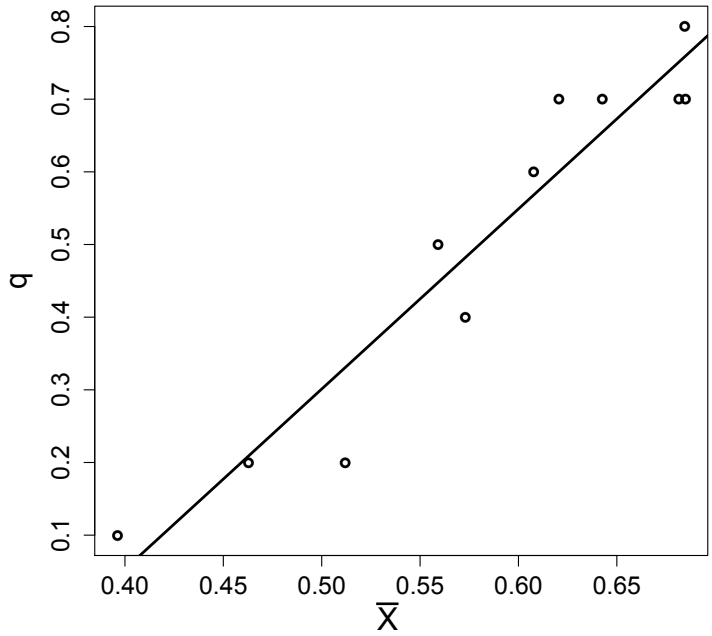


Figure 3.3: q estimation from the average of semantic similarity measures.

3.3.3.1. Top contexts

After the training phase, we applied the different steps of the methodology to participants' activity data extracted using the Facebook API. The resulting social contexts confirmed that (i) different users have different contexts and (ii) contexts referring to the same concept for different users are better defined using a tag cloud than a unique word. Table 3.1 shows the translation of the top topic terms in four different contexts of four different users. Although each of the two contexts is associated with two of the top generic interests (topics) detected in our study (*football* and *politics in Spain*), each context contains different terms

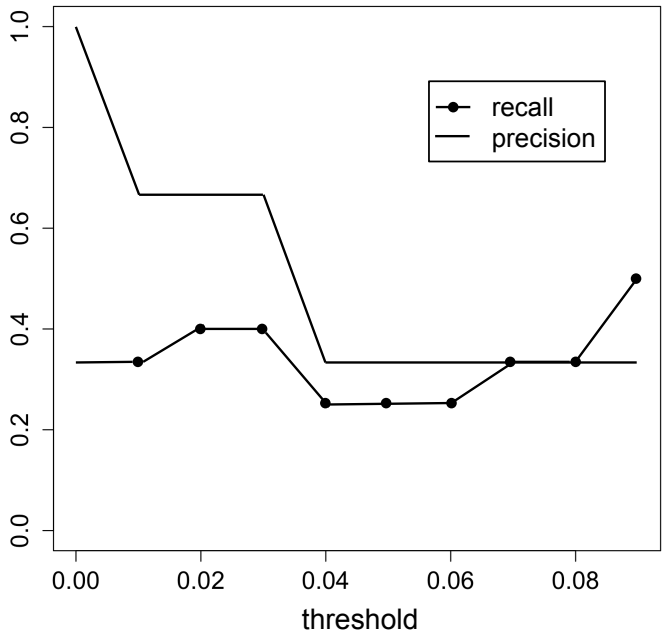


Figure 3.4: Recall and precision for different $Th_{context}$.

since they represent slightly different realities. For example, user 1 represents the view of an active political party member, whereas user 2 represents the view of a Spanish citizen and his complaints about the political, economic and social situation in Spain. In the case of the football topic, user 3 talks about professional football, whereas user 4 comments about the amateur football team in which his young son plays in the city of Vigo. Although contexts related to politics are found for most of the users, they are not suitable for the recommendation of any deal in our study. Meanwhile the football pool discount voucher is a good recommendation for football fans in general. However, in view of their tag clouds, it would be a good recommendation for user 3 – interested in professional football, but, a priori, not for user 4 – interested in amateur football. Note also that not having considered users' opinions (opinion mining is out of the scope of this dissertation) could cause this method to recommend ads to users who talk about a topic because they hate it. For instance, user 3 talks about *Real Madrid* because he dislikes it and, therefore, a possible discount on a Real Madrid T-shirt would not be a correct recommendation for him.

user 1	***user 2***	***user 3***	***user 4***
PSOE	vote	match	Celta
PP	unemployment	player	team
Rubalcaba	strike	football	match
president	trade	European	Vigo
government	minister	Barça	goal
campaign	cutback	Real Madrid	champion
election	crisis	referee	football
Rajoy	bailout	penalty	under-10
Bárcenas	tax	Messi	league
party	bank	goal	tournament

Table 3.1: Translation of top topic terms in four different contexts

3.3.4. Evaluation of users' satisfaction

After calculating participants' social contexts we asked them the following questions related to the contexts inferred:

1. *Point out the topics you usually talk about on Facebook.*
2. *We have identified that you talk about: (exemplars). Is this correct? Answer with a value between* 1 *and* 5 *(*1 *totally incorrect,* 5 *totally correct). Is there any other topic you regularly talk about on Facebook and we have missed? If so, indicate it.*
3. *For each series, underline the related words and add, at least, one term to each one.*
4. *Sort the next deals according to your preferences.*

Starting from the collected data, we tested (i) the concordance between the contexts identified by our method and the topics stated by the participants and (ii) the concordance between the optimal deals obtained by our method and those deals ranked by the users. The results are shown in the following sections.

3.3.4.1. Users' contexts inference

To evaluate participants' satisfaction with the social contexts inferred with our strategy we take into account both the contexts inferred and the answers of the participants to the first three questions of the questionnaire. Table 3.2 shows the comparison between human judgment and our method, where *h.j.* means *human judgment* and *o.m.*, *our method*.

	Average	***Std Dev***
Participants' satisfaction (over 5) -*h.j.*-	$2,933$	$0,742$
Number of topics -*h.j.*-	$3,967$	$2,465$
Number of clusters -*o.m.*-	$5,950$	$1,023$
Number of representative clusters -*o.m.*-	$3,333$	$1,867$
Number of tags by cluster -*o.m.*-	$18,414$	$17,768$
$\frac{\text{Related words -}h.j.\text{-}}{\text{Total words -}o.m.\text{-}}$ by cluster	$0,396$	$0,176$

Table 3.2: Comparison between human judgment and our method

The satisfaction of the participants with the exemplars of the clusters is not excessively high ($2,933$ over 5), which indicates that the centre – exemplar – of the clusters returned by *Affinity Propagation* is not representative for the cluster. However, topics pointed out by the participants are semantically close to the tags of at least one of the clusters returned by the algorithm. As an example, a participant in the experiment identified a topic with the name of her specific degree, which is clearly related with the exemplar *course*.

Regarding the *number of representative clusters* – contexts, participants stated different number of topics, which agrees with our premise that the number of contexts is different for different users, hence the importance of not fixing topics a priori. Also, our method identifies, in general, fewer contexts than the participants. After inspecting the original data, we detected that users stated topics about which they believe that they talk a lot but really they barely do, or include generic words to identify topics, making these topics difficult to recognise. For instance, a common topic stated by the participants was *daily routine*.

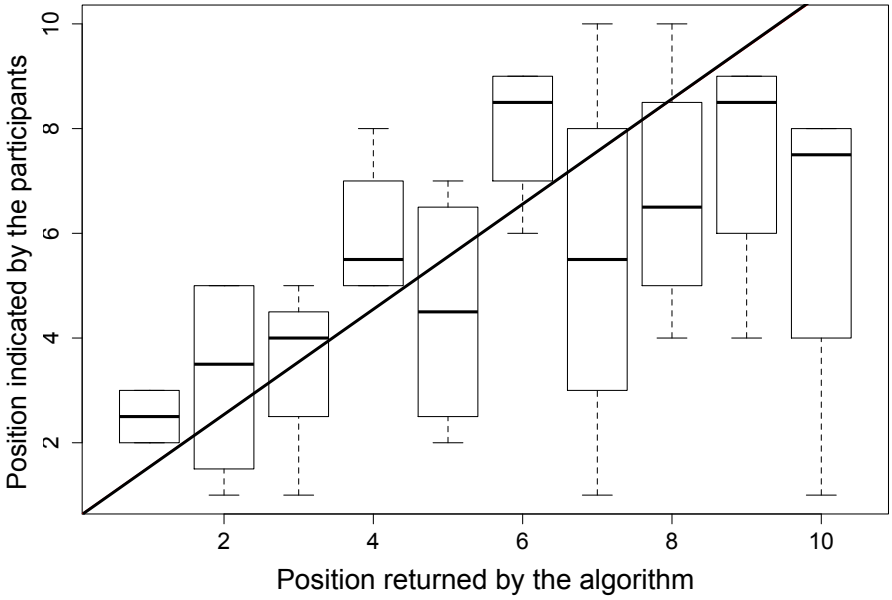


Figure 3.5: Concordance between deals rankings by users and deals rankings by our method.

The rate of related words by cluster identified by the participants – *(Related words)/(Total words) by cluster* in Table 3.2 – is not excessively high. Relations established are subjective since they depend on users' opinions. For instance, a user pointed out the relation between *Ramos* and *Ronaldo* (two *Real Madrid* footballers) within the social context whose exemplar was *Real Madrid*, but he did not point out the relation between them and *Cristiano* (another player of this team). Even taking this into account the rate is still low.

In fact, users are not as reliable as we expected when they identify the topics they talk about on Facebook since the answer to question 1 should be contained in both the wording and the answer of question 2, which does not happen in most cases.

3.3.4.2. Application of our method to social publicity

We consider the last question of the questionnaire: *Sort the next deals according to your preferences* (providing users with the ten deals extracted from *Groupon*) to evaluate the application of our method to social publicity. The

box-and-whisker diagram in Figure 3.5 represents the concordance between the position of the deals stated by the participants and that obtained with our model. The bottom and top of the box are the 25^{th} and 75^{th} percentiles, the band inside the box is the median and the ends of the whiskers are the 10^{th} and 90^{th} percentiles. The concordance between the preferences of the participants and the deals inferred by our method is higher in the top positions, the most important for our purposes. Concordance in low positions is worse since the cosine similarity is, in general, quite close between some deals and others. However, this is not a big problem for our social publicity application, since we are interested in the top positions of the ranking (those preferred by the users).

3.4. Experiment 2: comparison of clustering techniques on Twitter

One of the notable conclusions from the previous experiment is that users are not as reliable as desirable when they are asked about the topics of interest that they manifest in social media sites. This, together with the fact that most profiles on Facebook are private, which increases the difficulty of obtaining data and participants for a large scale evaluation of our methodology, led us to consider the use of a social site with more public data available. This led us to Twitter [twia], the popular microblogging service launched in 2006 [KGA08] that allows users to send messages – tweets – of up to 140 characters to people subscribed to their streams and where each user has a profile, designated as private or public, though most are public. Using Twitter as testbed, we devised a new experiment that takes advantage of the implicit mechanism to classify tweets in Twitter according to their subjects: *hashtags*[2]. However, the facts that (i) not all tweets are labeled with hashtags and (ii) one unique topic could be associated with different hashtags prevent us from using users' hashtags as the real users' topics against which to validate our results. Nevertheless, hashtags are still a highly valuable source of information about the content of tweets. For this reason, we propose

[2]The # symbol, called a hashtag, is used to mark keywords or topics in a tweet. It was created organically by Twitter users as a way to categorise messages. *https://support.twitter.com/articles/49309-using-hashtags-on-twitter#*

an experiment that takes advantage of hashtags to evaluate the performance of three different clustering algorithms – PAM, AP and UPGMA – when considering the structure of Wikipedia as an external source for assessing semantic closeness between words. Specifically, the experiment deals with building fictitious users' profiles by means of the aggregation of the tweets that contain specific hashtags. The next section explains the details of the experiment.

3.4.1. Problem definition

Starting by selecting a set of h hashtags (semantically distant enough), we collected the $1,500$ most recent tweets that contain each hashtag from the Twitter API and preprocessed them following the steps in Section 3.2. The set of hashtags selected for our study were *#apple*, *#realmadrid*, *#palestine*, *#sephora*, *#disease* and *#nba*. Under the premise that one *hashtag* deals with one unique topic, we ended with 6 different topics, each being represented by a group of tags together with their frequency of appearance (tag cloud). In order to build fictitious users' profiles, we merged the hashtags' tag clouds into global tag clouds. This resulted in 5 different fictitious users that *talked about* 2, 3, 4, 5 and 6 different topics respectively, these being *#apple* and *#realmadrid* for the first profile, *#apple*, *#realmadrid* and *#palestine* for the second profile and so one. Then, after calculating the semantic relatedness between their tags, we run the aforementioned clustering algorithms varying their input parameters and the threshold of relevance. Users' extracted topics are then represented by those clusters (whose weight is higher than the threshold) obtained when an unsupervised measure that considers both intra- and inter-cluster distances (the Silhouette width [Rou87]) is maximised. Finally, instead of gathering feedback by explicit user contribution, we use the implicit classifying method used by Twitter users to organise their tweets: *hashtags*. The next section briefly details the semantic relatedness measure and the clustering algorithms used in our experiment. Then, Section 3.4.3 describes the obtaining of the topic-clusters when the parameters are estimated by the maximisation of the Silhouette width. Finally, in Section 3.4.4 we use supervised measures (similarity between fictitious users' tag clouds and the tags in the resulting clusters) to validate the use of Silhouette width as estimator.

3.4.2. Particularising the model to this scenario

Our mechanism of interest extraction starts obtaining tweets by means of queries to the Twitter Rest API and extracting their relevant words by only considering lexical units that refer to fixed entities with meaning. To this aim, we used *Stanford CoreNLP* [MSB+14]. We filtered the text using POS tagging and lemmatisation (which identifies each word lemma, or citation form), only keeping nouns in their citation form. The resulting words form the set of input data points to the clustering algorithms.

Although hundreds of clustering algorithms have been proposed to date, we selected three of them to use in our experiment: (i) a traditional partitional clustering algorithm – *Partitioning Around Medoids (PAM)*, (ii) other partitional algorithm that, unlike *PAM*, does not need to indicate the number of resulting clusters – *Affinity Propagation (AP)* – and (iii) a classical hierarchical algorithm – *Unweighted Pair Group Method with Arithmetic Mean (UPGMA)* –.

Any clustering algorithm needs to know the similarity between data points which, in this case, is the semantic relatedness between the words to cluster. We used a measure based on the hyperlink structure of Wikipedia, *Wikipedia Link-based Measure (WLM)*, to assess this relatedness. See Section 3.2.3 for a detailed description of both the clustering algorithms and the semantic relatedness measure.

3.4.3. Parameters Estimation by Unsupervised Measures

Supervised measures are the best way to evaluate the performance of clustering algorithms in practical applications. Although in the case of our fictitious users we previously know their topics of interest, for the main goal of our research – extracting topics from users' tweets – topics are not known in advance, but they emerge from the clustering algorithm. For this reason, we used Silhouette width [Rou87], a clustering validation measure which indicates the strength of a cluster or how well an element was clustered, to evaluate the performance of the algorithms. The Silhouette width of a resulting cluster c, $Silh_c$, is calculated by:

$$Silh_c = \frac{(b_i - a_i)}{max(a_i, b_i)}, \tag{3.14}$$

where a_i is the average distance from the point i to all other points in i's cluster, and b_i is the minimum average distance from point i to all points in another cluster. We selected the variation of the Silhouette width (its average value over the resulting clusters) with the variation of the parameters of input to the algorithm (k, q), as the function to optimise, obtaining, in this way, the best selection of clusters from the input data. Table 3.3 represents the maximum values of Silhouette width obtained when different numbers of hashtags – different users' profiles – (from 2 to 6) are considered and the k – q – indexes at which these values are reached.

algorithms	***AP***		***PAM***		***UPGMA***	
|Hashtags|	$Silh(q)$	q	$Silh(k)$	k	$Silh(k)$	k
2	0.110	0.9	0.133	45	0.152	64
3	0.120	0.9	0.129	73	0.197	81
4	0.120	0.9	0.123	74	0.189	96
5	0.094	0.9	0.102	86	0.161	115
6	0.097	0.8	0.106	101	0.141	165

Table 3.3: Maximum values of Silhouette width.

The results clearly show a correlation between the number of hashtags considered (identified by topics) and the input parameter k both in the PAM and UPGMA algorithms. This correlation is not seen in the case of the input parameter q which controls the data points preferences in AP. With respect to the maximum values of $f(q)$ and $f(k)$, this correlation is not observed either. A comparison among the highest values of $f(x)$ of the three algorithms reveals that UPGMA is the one which produces the clusters with the highest quality, followed by PAM and, lastly, AP, which achieves the worst results.

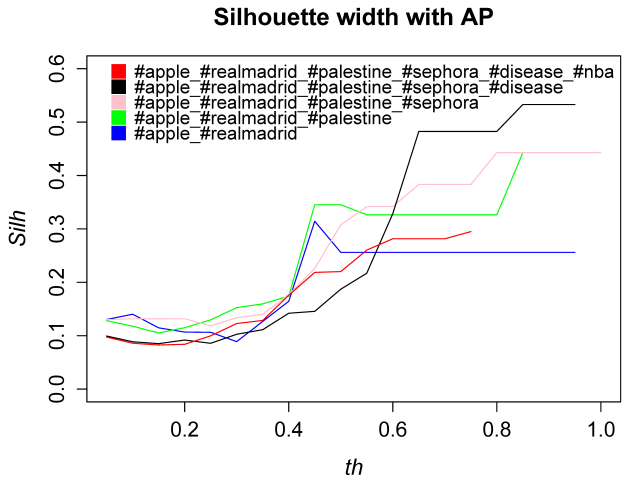
Once k – q – is fixed, not every resulting cluster matches with one hashtag, but the number of clusters is usually higher than the number of hashtags: there are words not semantically-related or weakly-related to the words in the set, which form clusters with few words or even with only one. In order to discard non-relevant clusters, we define the relevance of a cluster as the sum of the frequencies

of its tags. A cluster will be discarded if its relevance is lower than a threshold $Th_{context}$, which we define as $Th_{context} = th * w_{MAX}$, w_{MAX} being the sum of the frequency of the resulting cluster with the highest frequency and th an index between 0 and 1. To estimate the optimal value of th we again use the Silhouette width, but this time only considering the set of relevant clusters. Figure 3.6 shows the variation of the Silhouette width with the variation of th (i) for the different clustering algorithms and (ii) different number of hashtags (topics). Please, note that for higher values of th all clusters except one are discarded and the Silhouette width cannot be obtained.

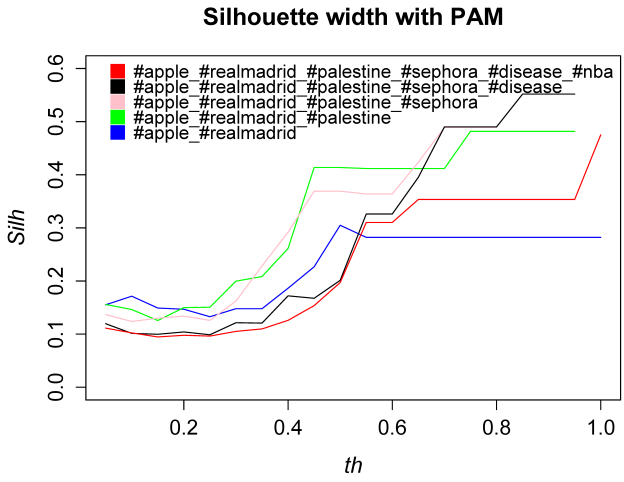
Clearly, the higher the threshold and, therefore, the fewer clusters that are considered, the higher the Silhouette width. But, this does not mean that we should discard clusters whose weight is not very close to the highest weight. The results show that the Silhouette value hardly varies with the incrementing of the threshold from 0.55 in UPGMA, 0.65 in AP and 0.7 in PAM, values that could be optimal for this parameter.

3.4.4. Evaluation Results by Supervised Measures

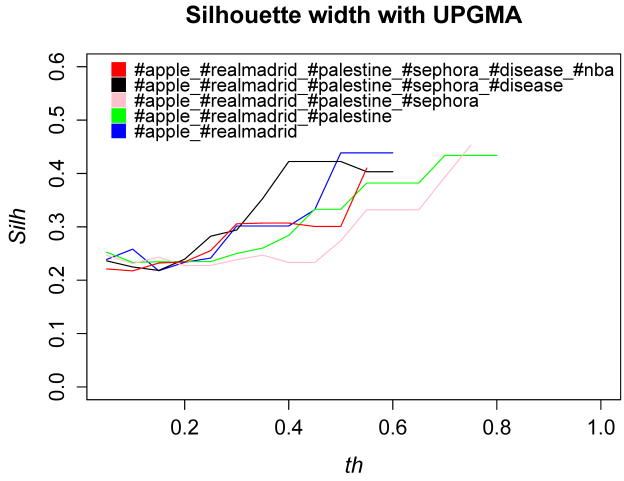
We use supervised measures, i.e. measures that consider external information about data – hashtags in this case –, to validate our topic extraction methodology and, specifically, the estimation of parameters by optimising the Silhouette width. Under the premise that the tweets that contain the same hashtag deal with the same topic, each of the h tag clouds should correspond to a resulting cluster. To this aim, we define the function $f(k)$ ($f(q)$), which represents the average of the similarity between each topic – hashtag – and its nearest cluster, defining this similarity as the difference between (i) the sum of the frequencies of the tags included both in the cluster and in the topic tag cloud and (ii) the sum of the frequencies of those tags included in the cluster but not in the topic tag cloud:



(a)



(b)



(c)

Figure 3.6: Silhouette width for different thresholds for (a) AP, (b) PAM and (c) UPGMA.

$$f(x) = \underset{i}{avg}\left\{\frac{\max_{j}\left\{\sum_{(t\in TC_i)\cap(t\in(C_j))} w_t - \sum_{(t\notin TC_i)\cap(t\in(C_j))} w_t\right\}}{\sum_{t\in TC_i} w_t}\right\} \tag{3.15}$$

where $x \in \{q, k\}$, TC_i represents the tag cloud of the i-topic, C_j the tag cloud of the j-cluster and w_t is the frequency of appearance of the tag t in the topic tag cloud. In view of Equation 3.15, the higher $f(x)$, the higher the correspondence between the clusters obtained and the h expected. So, the optimum value of the parameters (k – or q in the case of AP) is the one that maximises $f(x)$. Table 3.4 shows the maximum values of f when different numbers of hashtags (from 2 to 6) – different users' profiles – are considered and the k – q – indexes at which these values are reached.

algorithms	***AP***		***PAM***		***UPGMA***	
|**Hashtags**|	$f(q)$	q	$f(k)$	k	$f(k)$	k
2	0.166	0.85	0.242	5	0.394	5
3	0.168	0.05	0.241	17	0.240	39
4	0.144	0.05	0.194	23	0.201	43
5	0.178	0.05	0.204	27	0.198	95
6	0.134	0.85	0.141	80	0.160	153

Table 3.4: Maximum values of f(x).

In view of the results, the values of k and q when $f(x)$ is maximum differ from those obtained with Silhouette width. However, as in the case of the Silhouette, there is a correlation between the number of hashtags (topics) and k when $f(x)$ is maximum, which also means a correlation between k when f is maximum and k when the Silhouette width is maximum. A comparison among the highest values of $f(x)$ of the three algorithms reveals that AP is the one that produces the clusters more dissimilar from the original tag clouds, whereas PAM and UPGMA achieve similar results (except at the ends, where UPGMA behaves better).

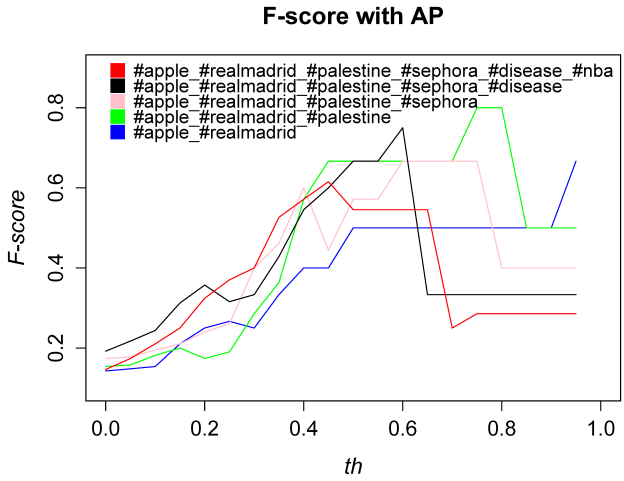
In order to validate the threshold which controls the relevance of a cluster, $Th_{context}$, we proceed in the same way as when using unsupervised measures, but

using F-score instead of Silhouette width. Figure 3.7 shows the variation of the F-score with the variation of th (i) for the different algorithms and (ii) different number of hashtags (topics). The results reveal that the number of hashtags and the value of the threshold for which the F-score is maximum are uncorrelated. Considering all the cases (results for different numbers of hashtags), the optimal value of th in UPGMA is between 0.5 and 0.55, while its value is slightly higher, approximately 0.6, in AP and 0.7 in PAM, values very close to those obtained taking into account the Silhouette width (0.55, 0.65 and 0.7 respectively).

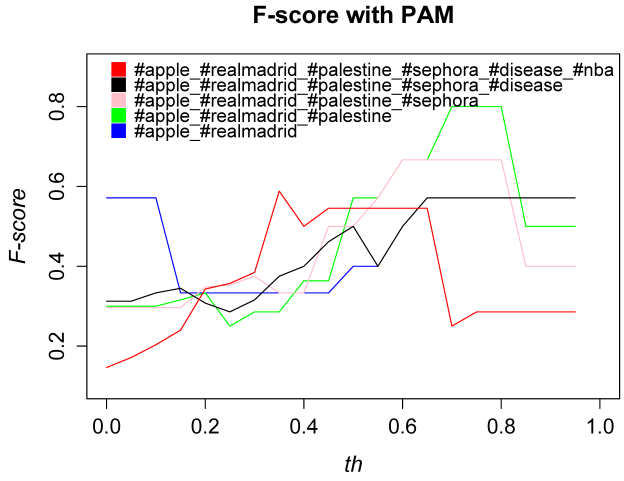
The estimation of the input parameters for the clustering algorithms and the threshold that determines the relevance of a cluster, both using supervised and unsupervised measures, reveal the existence of a correlation between the values of k, q and th estimated in one case and the other. Although the k parameters that maximise $f(k)$ and Silhouette width are different, the Pearson correlation between these ks is very high (0.987). With respect to the clustering algorithms (AP, PAM and UPGMA), the results show that UPGMA reaches the highest values both in $f(k)$ and in Silhouette.

3.5. Discussion

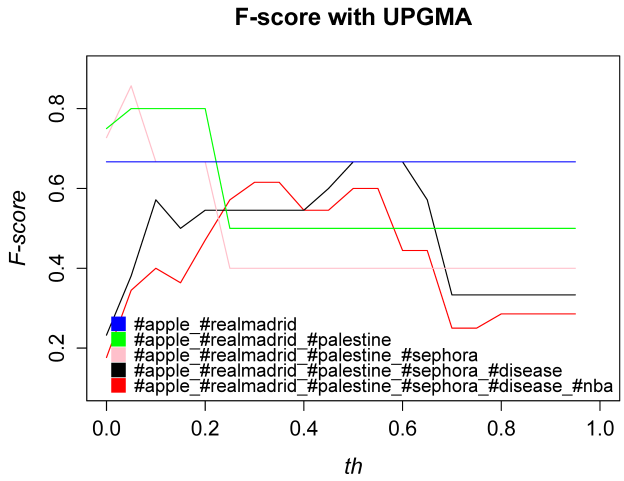
Two important issues have motivated this chapter: the importance of NLP-analysing user generated content in social media sites without any a priori categorisation of interests and the pursuit of a local strategy easily deployable and extensible. We provided a solution for the inference and management of users' social contexts that allows the personalisation/socially-enhancing of services as in the case of the *Groupon* scenario. A prototype of this scenario allowed us to conduct a user study to validate the suitability of our proposal. The high concordance between the ranking of deals of our participants and that obtained with our methodology, especially in the top positions, suggests the suitability of taking into account users' content spontaneously generated for this social publicity scenario. However, results when asking participants directly about their topics of interests presented several incongruences, which suggests that users are not reliable when they are asked broad questions regarding their interests. We also found that the bag of words representation allows distinguishing between interests



(a)



(b)



(c)

Figure 3.7: F-scores for different thresholds for (a) AP, (b) PAM and (c) UPGMA.

(contexts) that a priori, or using classification techniques, might be considered equal, such as in the two different contexts identified in our first experiment: "football" – amateur versus professional – and "politics in Spain" – from the view of a common citizen versus the perspective of a politician. This higher granularity in the representation of interests (contexts) implies higher flexibility in the definition and use of interests than with traditional LDA-based approaches ([HD10, ZYL+12, QAC12, DOMA12]).

In addition to our social publicity strategy, other services have already considered social media sites as advertising mediums; *Facebook Ads* [Facb] and *AdLemons* [adL] are good examples of this. But, based on existing open publications, we cannot be sure that they consider users' online lives, that is, the topics that users talk about, the content of the photos they share, etc. We can only be sure that the information that subscribers have consciously included in their profile is being taken into account, but nothing is known about the spontaneous content that they unconsciously generate and which emerges from their online lives. However, our user study demonstrated that this unconsciously generated information would be really useful for advertisers to accurately target potential customers.

Although from the Facebook API – with users' permissions – we get all data we need for extracting interests using our methodology, the fact that (i) getting participants willing to participate (for free) in a user study is an arduous task and (ii) the number of queries to the API in a time is limited, makes Facebook not the best medium for a large scale study. However, the public – open – character of Twitter allowed us to conduct this large scale study, even avoiding user-related issues, to validate – and compare individual steps of – our methodology. In fact, results from this study confirmed the suitability of using the Silhouette width for selecting the input parameters to the clustering algorithms and revealed that UPGMA is the best performing algorithm.

3.6. Summary

In support of the thesis that an intermediary model of the user constructed by properly mining user generated content in social media can be exploited to create

or improve technological social applications, we have proposed and evaluated in this chapter an algorithm to extract users' interests from textual publications that individuals freely post on social media sites. Our methodology, based on data mining and natural language processing techniques, can be applied without any a priori knowledge about the number and categories of interests. Also, the fact that these interests are represented by bags of words allows distinguishing between interests that a priori could be considered equal and might simplify the the personalisation of almost every application. We successfully evaluated this methodology by means of (i) a user study on Facebook and (ii) taking advantage of Twitter hashtags, the tool par excellence to mark keywords or topics in tweets. Results from the former study showed that the consideration of the content spontaneously generated by users allows the accurate prediction of user-preferred deals, converting our proposal into a good strategy for social publicity. But this study has also brought insights into how users are not reliable when they are asked about their interests. The study using Twitter that overcame the limitations of unreliability of participants and lack of data of the Facebook study, revealed (i) that a hierarchical clustering (UPGMA) is the best performing algorithm and (ii) the suitability of using Silhouette width for selecting the input parameters to the clustering algorithms. Apart from the social publicity scenario, knowledge of users' interests could be beneficially applied to recommender systems or social media dashboards, as detailed later in Chapter 5.

4

A user-centred measure to compute tie strength

In addition to expressing their interests, feelings and likes, social web technologies allow individuals to communicate with one another, developing social relationships. Although social media treats everybody the same, not all these relationships are created equal: we have from close friends to just acquaintances, and the rest of relationships fall everywhere along this spectrum. Thanks to the widespread use of these technologies, there is much information available about social ties between individuals, and the study of these embedded interaction social networks has been a recurring theme in research [GK09, WBS+09, BBK+11, KN09, XNR10, HRW09]. Although these studies obtain with great accuracy the interaction network underlying one social site, to date, little research has focused on measuring the perception that one user of several social media sites has about the strength of his relationships with others.

As an essential part of our proposed model of *social spheres*, in this chapter we present our user-centred measure of tie strength between two individuals. That is, we describe our methodology to assess the closeness that one user perceives of his relationship with another using evidence of their interaction activity in different online social sites – the tie strength between two users, u and v, from u's perspective. We show how we validated our measure by means of a user study on Facebook whose participants were asked to classify their contacts into groups of closeness, and how our measure was used to predict the group in which each contact was classified.

In relation to this thesis, our validation suggested that it is possible to measure the strength of the tie between one individual and any other social web user with great accuracy and with nothing other than his permission to retrieve evidence of interaction in the online services in which he has an account. This fact may be exploited to improve a great number of applications such as anti-spam e-mail filtering and group recommendation as indicated later in Chapter 5.

4.1. A model for user-centred tie strength calculation

A key contribution of this dissertation is a model to build users' social spheres using only interaction data retrieved from public APIs (with users' permission). In order to build these spheres, in our previous chapter we have explored the problem of extracting interests (social contexts) by applying NLP and clustering techniques to the data and metadata linked to users' interactions in social networks. Now, we describe our measure to assess the strength of the user's ties by using signs of interaction available from social sites' APIs (private messages, retweets, mentions, ...). To this aim, and contrary to previous approaches, we take into account different types of interaction, the time in which interactions occur, the people involved in them and the frequency of the interactions with the rest of the user's contacts.

The majority of previous proposals in assessing tie strength from users' data in

social sites (social networks, blogs, email, etc.) deal with inferring the interaction network that underlies the site. However, we are not interested in obtaining this interaction network, but our proposal is centred around the user – henceforth, the target user. Specifically, we aim to detect people with whom the target user usually interacts through social sites. As users do not interact with each other in the same way and with the same frequency, we develop a model to measure the closeness that one user perceives of his relationship with another from their interaction activity in online social sites (that is, the tie strength between two users, u and v, from user u's perspective). This subjective point of view may mean that the tie strength from user v's perspective is different, resulting in asymmetric tie strength. For instance, if user u often chats with v but also with other users whereas v only chats with u, their tie from v's perspective will be stronger than from u's. Moreover, note that the tie strength from u's view (v's view) will depend on not only their interactions, but also the interaction between u (v) and the rest of the online users with whom he interacts. So, although the level of relationship between them does not vary, it is possible that the tie strength between them from the target user's view is different due to changes in other relationships. For instance, in the previous example, if u keeps his chatting patterns whereas v now chats with more friends, their tie from v's perspective will be weaker than in the previous situation.

Relationships are formed by repetitive behaviours or reciprocal actions between individuals when they present certain persistence [Nad57]. Bringing this to the online world leads us to consider interactions between individuals as signs of the existence of a relationship or tie between them, for which, henceforth referred to as "tie signs". In order to compute the strength of this relationship or tie strength we take into account these manifested interactions in social media. As a result, our model provides indexes, with values from 0 to 1, that represent, from the target user's view, the tie strength with each one of the individuals in his *social sphere*.

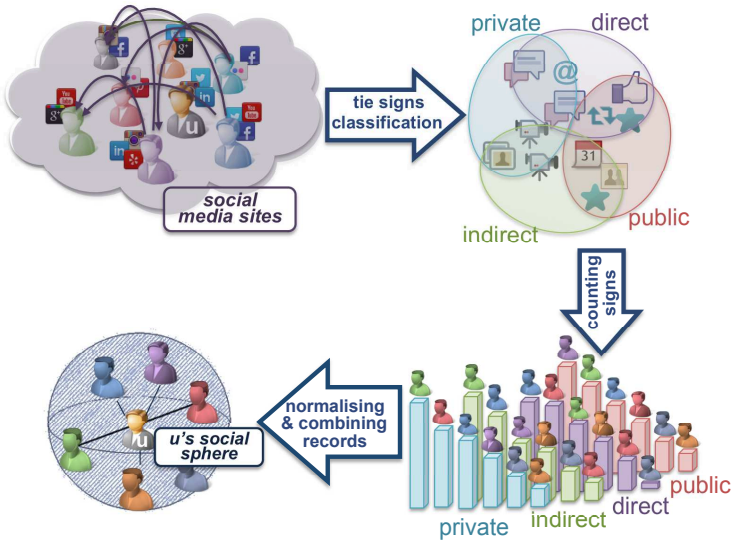


Figure 4.1: Inferring a user's social sphere

4.1.1. Tie strength calculation

Figure 4.1 shows an overview of the steps to obtain users' social spheres, i.e. computing the tie strength with their contacts, from evidence of interactions in social media sites. The process starts by monitoring users' activity on these sites making use of their public APIs, always with users' permission. From this monitoring process we get only users' activity that implies any kind of interaction between them (tie signs). Tie signs are, for example, exchanging private messages, being tagged in the same photo, attendance at the same event, etc. In the next step, these tie signs are classified regarding to their type – nature – into *private*, *public*, *direct* or *indirect*. Tie signs are classified as *private* or *public* depending on whether they occur between close friends or between simple acquaintances; they are classified as *direct* or *indirect* depending on whether they imply explicit communication or common interests. This classification is necessary since not all tie signs have the same impact on the strength of the relationship. For instance, the fact that two users exchange a wall-post on Facebook has more impact on their tie strength than, for example, the fact that they belong to the same public group, since the former tends to occur between close friends, whereas the latter may occur between simple acquaintances or even strangers.

In order to represent the strength of a tie between two users u and v (from u's

view), we define an index, *tie strength index* (denoted by $T_u(v)$), whose value will be close to 0 for a weak tie and close to 1 for one strong. To obtain this index, it is necessary to (i) count the tie signs of each type between the user u and v and (ii) normalise these tie signs counts with respect to the counts of the rest of u's contacts. Then, the resulting values are combined by means of a weighted addition, to obtain their tie strength index. The set of tie strength indexes between u and all his contacts will finally form u's *social sphere*. Mathematically:

$$TS_u(v) = \sum_{k=1}^{N_k} \alpha_k \cdot f(|S_{u|k}(v)|) \tag{4.1}$$

where $\sum_{k=1}^{N_k} \alpha_k = 1$, α_k denotes the weight of the signs of the k-type in the user's *social sphere*, N_k is the number of different tie signs types considered (4 in the case of the proposed classification – *public*, *private*, *direct*, *indirect*), $S_{u|k}(v)$ is the set of tie signs of the k-type associated with the shared link between u and the user v and f is a normalisation function:

$$f(x) = \begin{cases} 0 & \text{if } \ 0 \leq x \leq \dfrac{\bar{x}^2}{x_{max}} \\ \dfrac{\ln(\frac{x_{max}}{\bar{x}^2}x)}{\ln(\frac{{x_{max}}^2}{\bar{x}^2})} & \text{if } \ \dfrac{\bar{x}^2}{x_{max}} < x \end{cases} \tag{4.2}$$

$\overline{x}$ and x_{max} being the mean and maximum value, respectively. So, $f(x)$ is close to 1 if $x > \overline{x}$, close to 0 if $x < \overline{x}$ and, finally, close to 0.5 if $x \sim \overline{x}$.

4.1.1.1. Relevance and gradual forgetting

As life itself, tie strength should be a dynamic index reflecting that old interactions are progressively less important and, for that reason, should have less relevance in the index calculation. Additionally, signs' relevance vanishes as the number of participants increases. For instance, being tagged together in a photo with five people should be more relevant than being tagged together in a photo with twenty people; at least, it may be assumed that in the first case the situation entails more closeness. Given that signs' relevance decreases (i) as time goes by

and (i) as the number of participants increases, we propose to adjust the weight of each specific interaction in the tie strength calculation applying the following decreasing functions to take into account the time at which the interaction occurs and the people involved in it:

$$d(\mu, n) = e^{-\mu \cdot n} \qquad d(\mu, t) = e^{-\mu \cdot t} \tag{4.3}$$

where n is the number of participants in the signs and t is the time since the latest update of the sign. The parameter μ represents the strength of the slope, i.e. the rate of *vanishing* of signs' relevance, taking the values $\mu_{r|k}$ for the participants and $\mu_{t|k}$ for time. Note also that the respective slopes depend on the type or nature of the sign (*public*, *private*, *direct* or *indirect*).

4.2. Tie Signs in social media sites

Social media users interact with each other using the facilities that social sites offer them. For instance, Facebook enables users to publish posts on their wall (or on other users' walls) or upload photos and/or tag them. However, on Twitter, for example, the wall does not exist, but when a user wants to connect with another, apart from using private messages, he can mention him in a tweet. Also, although Twitter enables the user to upload pictures (stored in external servers), they are seldom tagged. So, although different social sites provide their subscribers with different technical features to interact, we may find similar interaction facilities among them or, at least, used for the same purposes. Our model is aware of this, so that it defines a classification of evidence of interaction – tie signs – by type that is independent of the social site from which the tie signs were retrieved.

In this section we provide examples of Facebook and Twitter features that we consider evidence of interaction between users (tie signs) and their classification into the four types considered in our model (*public*, *private*, *direct* and *indirect*). We selected Facebook and Twitter to put the model into practice because they are two of the most well-known and most used social sites and because they pro-

Signs ($S_{u\|k}(v)$)	***direct***	***indirect***	***public***	***private***
Wall-posts on friend's wall	x			x
Private messages exchanged	x			x
Comments on friend's objects	x		x	
Comments on the same objects		x	x	
Likes of friend's objects	x		x	
Likes of the same objects		x	x	
Being tagged in the same photos or videos		x		x
Belonging to the same private group		x		x
Belonging to the same public group		x	x	
Attending the same private event		x		x
Attending the same public event		x	x	
Being subscribed to the same user		x	x	
Being subscribed to by the same user		x	x	

Table 4.1: Tie signs classification on Facebook

vide developers with public APIs to retrieve users' data (with their permission). However, although we only indicate evidence of relationship in Facebook and Twitter, the classification proposed in our model, and consequently the proper model, is general enough to be extrapolated to other social sites.

4.2.1. Tie Signs: the Facebook case

Facebook provides its users with the typical interpersonal communication features, where the *wall* is its highlight. Subscribers use the wall to post photos, videos, links and messages that may be enriched with any friends' comments. In addition, *mini-feeds* provide detailed logs of each subscriber's actions, so any friend may see at a glance his evolution on Facebook over time. As in any social network, security is a key factor and Facebook allows its subscribers to personalise the privacy settings to restrict access to the profile information, mini-feed, wall posts, photos, comments, etc. (only to friends, friends-of-friends, lists of friends, no one or all).

After a detailed analysis of Facebook features, how users interact and communicate and the data available through the API, we have identified the interaction signs whose classification by type is shown in Table 4.1. We include wall-posts and private messages in the direct category since they are interactions that take place at a specific moment to communicate something between two or more users. However, a user joins a group or attends an event because he is interested in the topic of the group or event, receiving information from other members of the group or event in an indirect way. Analogously, relationships of subscription are included in the indirect category. In addition, we also include "Likes" and comments in the direct category when the user u comments on – likes – one user v's object and in the indirect category when both users comment on – like – the same object. Being tagged in the same photo or video is considered an indirect sign since usually we do not know if the users communicated at the time when the photo was taken or the video was recorded.

We consider wall-posts, private messages, being tagged in the same photo or video, membership of the same private group and attendance at the same private event to be included in a user's private range, whereas comments, "Likes", membership of the same public group, attendance at the same public event and relationships of subscription to belong to a user's public range. The reason is that the former tends to happen between close friends while the latter may also happen between users who are simply acquaintances.

4.2.2. Tie Signs: the Twitter case

In the web-based microblogging service Twitter users can link to ("follow") others and see their tweets, but it is not necessary reciprocated ("be followed"). An important feature of Twitter that users see when they log in is the *home_timeline*, a collected stream of Tweets posted by the user and the users he follows listed in real-time order. By default and norm, users' profiles and tweet streams are public but they may be made private.

Twitter users, apart from posting indirect tweets, can post direct ones; i.e. they can update with text addressed to everyone or directly addressed to a specific user in which case the sign "@" followed by the username of this specific

Signs ($S_{u\|k}(v)$)	***direct***	***indirect***	***public***	***private***
Mentions (replies)	x			x
Private messages exchanged	x			x
Retweets friend's tweets	x		x	
Retweets the same tweets		x	x	
Marking friend's tweets as favourite	x		x	
Marking the same tweets as favourite		x	x	
Taking part in the private same list		x		x
Taking part in the same public list		x	x	
Sharing the same Hashtag		x	x	
Common Followers		x	x	
Common Followees		x	x	

Table 4.2: Tie signs classification on Twitter

user (*@username*) is included in the tweet. *@username* can also be used to refer to someone. However, Honey and Herring [HH09] prove that the most common use of "@" is to indicate addressivity (in more than 90% of cases) followed by reference use (mentioning someone). Another important Twitter feature is the use of "#" (*hashtags*) to mark keywords or topics in tweets, making tweet classification easier. Also, as in many social sites, private messages are exchanged between users to share confidential information.

After a detailed analysis of Twitter features, how users interact and communicate and the data available through the API, we have identified the interaction signs whose classification by type is shown in Table 4.2. We consider mentions and private messages to be in the direct category since they are interactions that take place at a specific moment to communicate something between two or more users, whereas the fact that users take part in the same lists, use the same hashtags and follow common users (or are followed by the same users) are included in the indirect category, since they are related through common interests, receiving information from the rest of the users in the list, users who use the same hashtag and users who follow/are followed in an indirect way. In addition, we also include favourites and retweets in the direct category when user u retweets user v's

tweet or when u tags v's tweet as a favourite whereas they belong to the indirect category when both u and v retweet the same tweet or tag it as a favourite.

Mentions, private messages and taking part in the same private list fall into the private signs category, whereas retweets, favourites, taking part in the same public list, using the same hashtag and following relationships belong to the public signs category. Mentions are included in the private range because, although everybody may see the tweet in which they are included, this is addressed to the Twitter user who is mentioned; retweets fall into the public category because the user who posts the retweet wants to share with everybody an interesting tweet of another user. We also include favourites in the public category since they are similar to retweets, in the sense that they express the fact that the user likes the tweet. Finally, using the same hashtag belongs to the public type since hashtags are not private, but they can be used by everyone.

4.3. Validation

The evaluation of our model requires permission to access users' data and not all of them are willing to grant it. Nevertheless, we have developed an application to (i) extract users' interaction data on Facebook (with their permission) and, at the same time, (ii) ask users about their relationship with others, which allowed us to validate our model of inference against human judgement. This section reports the experimental evaluation of our model to infer the strength of the ties between social users using the aforementioned application. Note that the contextual dimension of social spheres is included in the previous chapter, and is out of the scope of this evaluation.

4.3.1. Problem definition

Our tests involved 22 (under)graduate students from the University of Vigo, their friends and relatives. We ended up with a diverse audience, with disparate demographic data and educational backgrounds, including nearly as many men as women with ages ranging from 22 to 51 and with different use of Facebook: from

users that interact with their Facebook friends several times per day to those who only use Facebook features once per week or even less. We asked participants to (i) let our application access their Facebook interaction data on their behalf and (ii) classify 30 of their Facebook friends (randomly selected) into four different groups of closeness (from close friends to simply acquaintances) in accordance with the perception that they have about their interaction level on Facebook. We selected four groups of closeness in accordance with the different intimacy levels defined in Dunbar's social brain hypothesis [Dun98] (*support clique*, *sympathy group*, *affinity group* and *active network*). The percentage (in average) of participants' friends included in these groups was 38% in the lowest closeness group, 26% in the next, 20% in the third group and 16% in the highest closeness group. Please, note that the distribution does not correspond exactly to Dunbar's hypothesis, since we only consider, for this study, (i) friends with whom participants have interacted at least once in the last year and (ii) 30 of their randomly selected friends. In addition, we asked them about how they thought that others would assess the relationships (equal, better or worse than them) – control question.

4.3.2. Fixing parameters for relevance and gradual forgetting

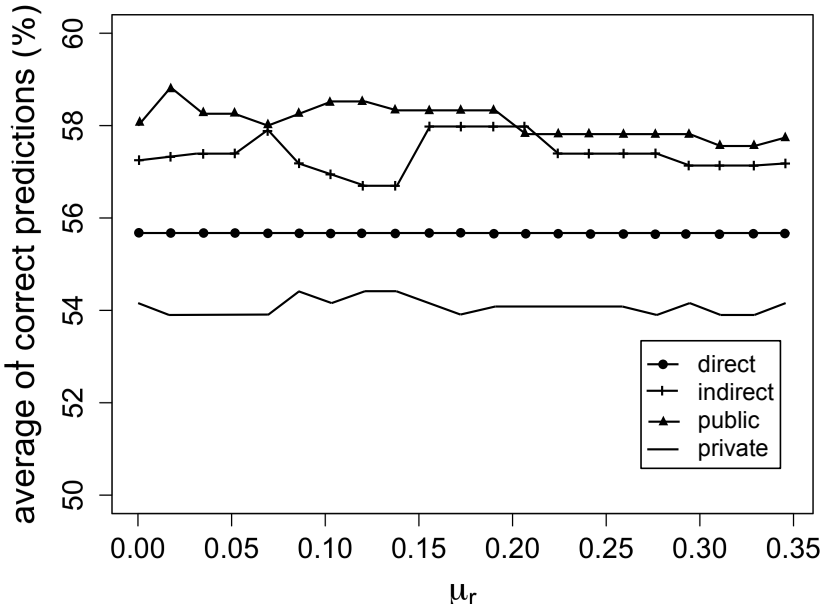
Once participants granted permission to our application, we retrieved their tie signs (those indicated in Table 4.1 without, due to privacy issues, private messages and secret groups) of the last year. We also calculated, applying the steps indicated in Section 4.1.1, the tie strength index with each of their contacts. In order to fix the parameters that control the importance of the relevance (μ_r) and the gradual forgetting (μ_t) in the tie strength calculation, we used multinomial logistic regression. Specifically, starting from the collected data, we predicted the groups of closeness in which each participant's friends would be included by a multinomial logistic model using (i) the groups indicated by the participant as the *response* and (ii) the tie strength values between the participant and their randomly selected friends as *predictors*.

We performed different predictions varying the parameters μ_r and μ_t involved in the tie strength calculation in order to find suitable values according to human

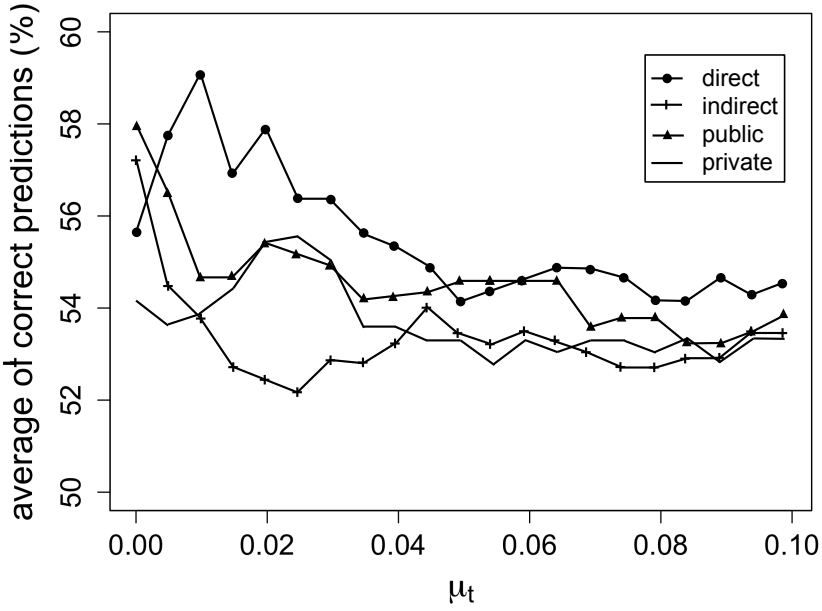
judgment. Please note that μ_r goes from 0, when the tie signs' relevance is not taken into account, to $\mu_r = 0.35$, when the importance of the sign drops to half of its original value (1) if only two people are involved. In the case of gradual forgetting the maximum value of μ_t considered is $\mu_t = 0.10$, which means that the importance of the sign drops to half of its original value when a week has passed since the sign occurred. Figure 4.2 shows the percentage (on average) of participants' friends that were correctly classified by the predictor with respect to the total number of friends to classify (30 in this case) for different values of (i) importance of the signs' relevance (μ_r) and (ii) gradual forgetting (μ_t).

The results show that, in general, the percentage of correct values predicted is close to 60% in many of the cases. Specifically, when only the relevance of the direct signs is taken into account, this percentage barely varies for different values of μ_r, but variations are significant in the cases of public and indirect signs. The reason is that usually in direct signs (such as the publication of posts on friends' walls) there are only two people involved in the interaction (the participant and the friend). However, in the case of indirect signs (such as membership in a common group) there are more people involved in the interaction, which allows the detection of differences. With all of this, the suitable value of μ_r that would increase the number of friends predicted correctly would be around 0.16 and 0.19 for all the four types of signs, except in the case of private signs, which would be a bit lower. Also, when considering gradual forgetting, the percentages of participants' friends correctly predicted vary with variations in the parameter that controls the importance of time in the interactions for all four types of signs. In this case, the lower values of μ_t are the ones that get the highest percentage, which means that participants need, on average, a long time to forget interactions with their friends. So, the optimum values of μ_t would be around 0 and 0.01, except again for private signs, whose optimum value would be around 0.02. With all of this, we fixed $\mu_r = 0.16$ and $\mu_t = 0.01$ as the optimum values for all types of signs (except for private signs where $\mu_r = 0.12$ and $\mu_t = 0.02$) to be used in the next step of the evaluation.

The percentage of correct predictions is also suitable for deciding the importance of the different types of tie signs (α_k) in the tie strength calculation. In this case, when both relevance and gradual forgetting are taken into account, the type



(a)



(b)

Figure 4.2: Average of correct predictions using multinomial logistic regression (%), by varying (a) the relevance (μ_r) and (b) the gradual forgetting (μ_t) parameters.

of tie sign with the worst results achieved is the private one. When public signs are considered, the percentage of correct predictions is one of the highest, both when relevance and time are taken into account. With respect to direct signs, the correct predictions are the highest when gradual forgetting is considered, but their results are not as satisfactory when relevance is taken into account. Finally, indirect signs achieve, in general, satisfactory results. Taking these results as evidence of predictive power, we fixed the weight distribution among the different types of signs (α_k) with $\alpha_d = 0.35$, $\alpha_i = 0.25$, $\alpha_p = 0.30$ and $\alpha_s = 0.10$ (for direct, indirect, public and private signs respectively). Please note that having taken into account neither private messages nor private groups limited the number of private signs considered in the calculation. Probably, if they had been used for calculation, correlation with human judgement would have been higher.

4.3.3. Experimental Results

Having fixed the values of the different parameters involved in the model, we calculate the tie strength between participants and friends taking into account the importance of both the number of people involved in the signs and the time since they occurred. Figure 4.3 shows the distribution of tie strength values between participants and their Facebook friends with respect to the groups in which participants included them. The median of the tie strength values increases with the closeness indicated by human judgement. We have used the same value for μ_r and μ_t (and also for α_k) for all participants. However, we are strongly convinced that users appreciate the effects of relevance and gradual forgetting in different ways. That is, a user may not consider his relationship with a friend to have weakened if they have gone a month without talking, but another user may consider a month without talking to be evidence that the relationship has weakened. So, we believe that training the parameters independently for different users we would have achieved better results. However, as training the system for every user is not feasible in a practical case we decided to consider a global model for this evaluation.

From Figure 4.3, we can also see that, although the median values of tie strength in each group are sufficiently separated from the median in the rest of the

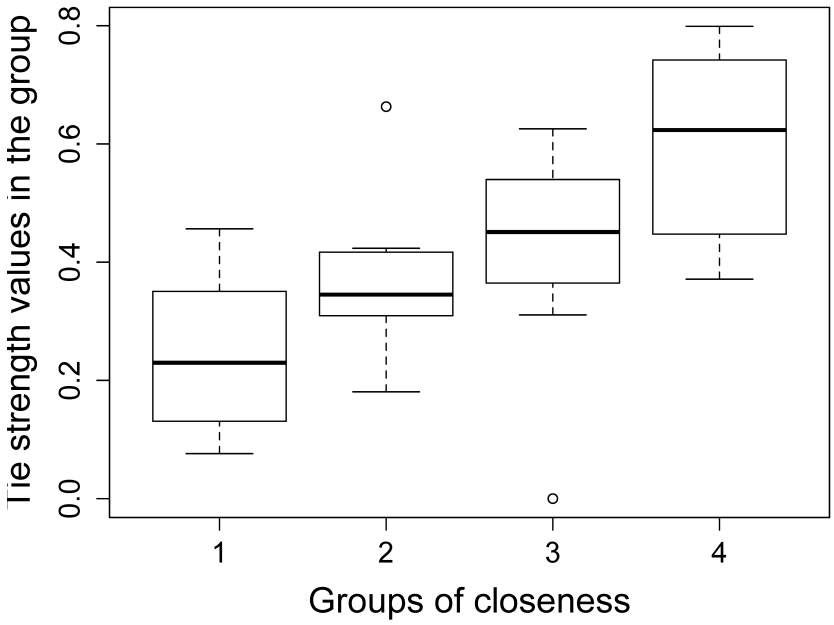


Figure 4.3: Tie strength distribution in each group.

groups, there is overlap between the 25th and 75th percentile across consecutive groups. This, together with the presence of outliers, explains the percentage of friends correctly predicted not being as high as desirable.

Finally, the control question to assess the reliability of the users when they are asked about their relationships on Facebook revealed that users are not as reliable as expected. For instance, some participants indicated that others would have assessed their relation with a given friend as being closer than them when they had given the relationship the highest mark, including the given friend in the closest group. An examination of the interaction data revealed that some participants included friends with whom they barely interact in the closest group.

4.4. Discussion

An essential part of our user-centred model to personalise applications is the tie strength assessment between users from their interactions in social sites. Even

though this is not the only measure proposed with this aim (see Section 2.2 for a review), it aims to improve previous ones (i) by taking into account key aspects of the interactions such as type, timing and people involved in them, as well as (ii) by using only information available through public APIs of social sites with users' permission. Using data from as many social sites as possible avoids losing information in the tie strength calculation process, since users usually relate through multiple social media platforms. In addition, it is possible that they do not interact with their contacts in all of them or with the same frequency. Contrary to previous approaches that compute tie strength at a network level [WBS+09, VMCG09, BBK+11], we provide a user-centred approach that models the perception that one individual has of his relations with others, and which is dependent on his relations with the rest of his contacts. Also, the perception of a tie between two individuals can, and will be different depending on whose view is considered.

In order to validate and refine our model, we developed an application to extract users' interaction data and discover users' opinions about their relationships on Facebook. Following Dunbar's social brain hypothesis [Dun98], we asked participants to classify their Facebook contacts into four groups according to their closeness, which we tried to predict by means of a multinomial logistic model. Results revealed that taking into account time and relevance of the interactions increases the number of friends that our predictor classifies correctly. With all of this, the percentage of relationships correctly classified was not very high (close to 60%). One possible reason could be that when users are asked about their relationships with others, they are unable to separate their *online* from their *offline* relationships, indicating strong closeness with some users that they barely interact with on Facebook (but they often do in real life) and weak with users that interact a lot on Facebook but barely in real life, which seems to be confirmed by the control question used in the experiment. Moreover, although participants were asked to classify a random sample of 30 contacts in the circles of closeness extracted from those users that they have interacted with at least once during the previous year, most of users' contacts did not fall into this category. This is in keeping with the work in [WBS+09], where Wilson et al. point out that, for most Facebook users, the vast majority of interactions occur only across a small subset of their social links.

Finally, relationships between users may exist in some contexts but not in others, depending on their interactions in the considered context. Consequently, following White's theory of *Netdoms* [Whi08] and considering interactions that happen in each context separately, our model may provide different users' social spheres depending on the context considered. That is, apart from taking into account the strength of their ties, taking into account the context of their lives in which their interactions happen. With regards to this, in our previous chapter we have explored the problem of extracting users' contexts by applying NLP and clustering techniques to the data and metadata linked to users' interactions in social media sites. Measuring the similarity between textual content linked to tie signs and the contexts of the user we would be able to classify tie signs into contexts (similarly to when classifying contacts into contexts in Chapter 3). Thus, considering only the signs that belong to one context we could calculate the tie strength indexes between the user and his contacts in the given context needed to build the *contextualised social sphere*.

4.5. Summary

In relation to the thesis that an intermediary model of the user constructed by properly mining user generated content in social media can be exploited to create or improve technological social applications, we have proposed and evaluated in this chapter an algorithm to measure the strength of the relationship (or tie strength) between two individuals from evidence of their interactions in social media sites. Although we only focused on evidence of relationships or ties in the social network Facebook and the microblogging service Twitter, the algorithm proposed here may be easily extended to other social media sites. In order to evaluate the proposed measure, we designed and implemented a Facebook application to obtain users' interaction data needed by our algorithm and to ask users about their relationships with their contacts. Our validation against human judgement revealed that (i) our measure produces an acceptable classification of individuals' contacts into circles of closeness and (ii) our control question confirmed that users are not as reliable as expected when they are asked about their Facebook relationships. In terms of possible applications of these results, and

especially of our tie strength measure, services such as recommender systems or e-mail readers might be socially-enhanced. Additionally, this tie strength measure in association with the measure to extract social contexts explained in the previous chapter might be used to personalise other services described later in Section 5.5.

5

Applications and pilot experiences

Having detailed the data mining techniques used to build the social spheres, both in terms of social contexts and tie strength, in this chapter we focus on their practical applications. We describe how our social spheres model could be easily implemented as an online service following the software-as-a-service paradigm. This service would be in charge of building, managing and delivering the spheres on request and always with users' permission.

In addition to the social publicity application described in Chapter 3 and used to validate our methodology for interest extraction, other applications may be socially-enhanced – personalised in a social context – taking into account our social spheres model. In this chapter, we also analyse the possibility of using the spheres in two socially-enhanced applications in charge of helping users (i) gain attention and (ii) find trustworthy experts in social media. For the former, we detail our experiment to study the relation between the diversity of topics talked about by users in their publications and the size of their audience. The

discovery of such a relation demonstrates the suitability of using social spheres – social contexts – in an application that alerts users when they should diversify or specialise the topics in their publication in order to increase their audience. For the latter, and in order to overcome users' isolation caused by their limited view of the system – they are only aware of what happens in their social spheres – we detail our social P2P proposal that takes advantage of social spheres – both in terms of tie strength and contexts – to help people find appropriate information or services more effectively. As these are not the only services to socially-enhance, we end the chapter by discussing other applications that might be personalised using social spheres.

In relation to our thesis, the results of this chapter show how our social spheres model can be used to create and enhance existing social technological services. Additionally, implementing the social spheres model as a service that follows the software-as-a-service paradigm and provides support for other services through a REST API increases the chances of being considered by other applications as the external tool for providing their users with satisfactory personalised experiences and without detriment to privacy.

5.1. Why use a social spheres service?

The main idea behind this dissertation is to provide a service-oriented solution for personalising applications using only user generated content in social media. To this aim, we envision a software component – *as a service* – that may be easily encapsulated in a great variety of socially-enhanced services: those whose behaviour is always aware of the user, and specifically, of his social component (contacts, contexts – interests – and strength of ties). This intermediary service, *mySocialSphere* in Figure 5.1, would work as a crawler with a double goal: (1) monitoring users' activity in online social sites to build their social spheres in different contexts of their lives and (2) providing other services with these social spheres to improve their effectiveness, releasing them from mining their users' social data. One notable aspect of our social spheres is that they are not restricted to contacts subscribed to the same online social site: *mySocialSphere* handles relationships although their interactions only occur in one social web site. For

instance, in Figure 5.1, the only common social site that the user u and the user v use and in which they interact is Facebook and, even so, *mySocialSphere* may detect properly their relationship and reflect it in their social spheres.

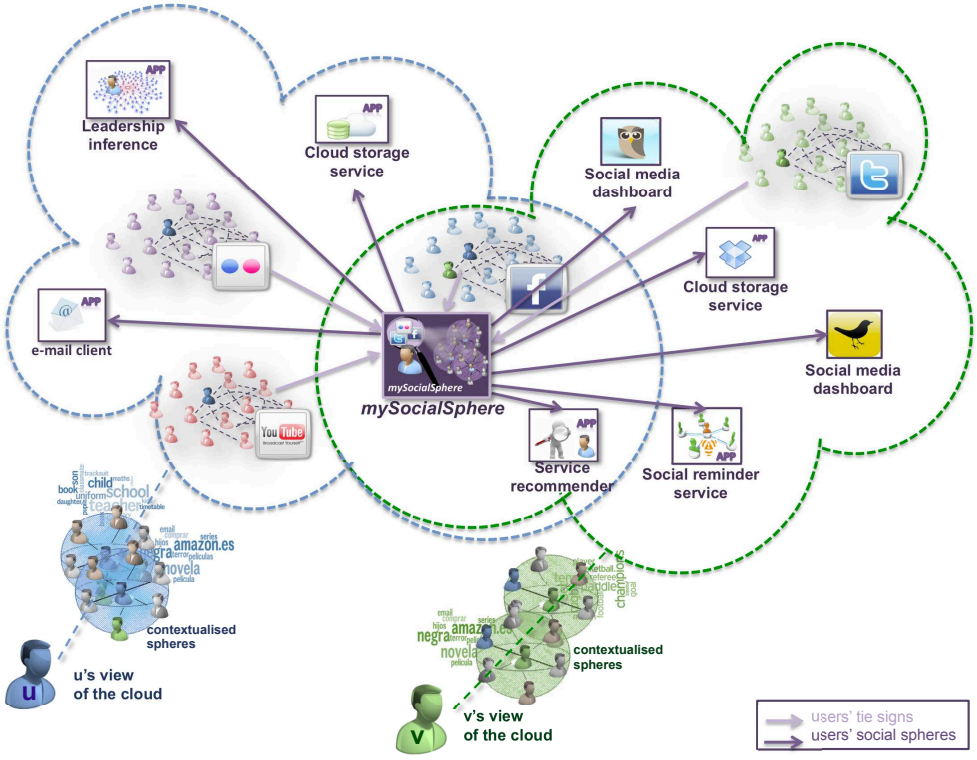


Figure 5.1: *mySocialSphere* Service and other socially-enhanced services

Figure 5.1 also contains some service-oriented services or applications that might be socially-enhanced or created using the social spheres provided by *mySocialSphere*. In this scenario, *mySocialService* gives users flexibility to change from one service to another when both implement a concrete functionality. That is, a user may use a social media dashboard (platform used to manage users' updates in different social networks) today and tomorrow change to another without losing the benefits of applying social spheres to enhance the new social media dashboard's work. Also, and although *mySocialSphere* is independent of the social site, some services, for instance recommenders, may need, apart from social spheres, other users' data (profiles, historical rating, etc.) to properly operate. However, others, such as the aforementioned social media dashboards, work properly knowing only the social spheres.

Finally, special attention should be paid to privacy concerns. Many users

underestimate the risks and potential threats related to information privacy in online social networks [Hug11], with adults tending to be more concerned than either adolescents or young people. Although *mySocialSphere* may need specific permissions to access private data – depending on the privacy configuration of the social media site under consideration – users are expected to be willing to grant them if socially-enhanced applications offer them sufficient added value. It is notable that, as was noted by Tan et al. [TQKH12], privacy concerns do not seem to directly affect users' acceptance of social networking web sites.

5.2. An architecture for Social Spheres

As mentioned, and with the aim of putting into practice our social spheres model, we propose an intermediary service, *mySocialSphere* in Figure 5.2, that provides two main functionalities for a specific user u: the construction of his social spheres – and contexts – from user generated content gathered from social media and the provision of these spheres to other services in order to be socially-enhanced or personalised. Below, we indicate some feasible technological solutions for each functionality, although one should keep in mind that other technologies might also be suitable.

Regarding the construction of u's social spheres, *mySocialSphere* would monitor users' activity in the Social Web (online social networks, blogs, wikis, etc.) and, following the model described in Section 1.1, update u's ties, their strengths and social contexts. Contrary to how it may seem, we do not aim to turn *mySocialSphere* into a social network with control over users' data. We strongly believe that users are the only owners of their data and should be able to store their spheres and contexts in whichever place they trust. For this trustable scenario to be possible, *mySocialSphere* could work with an XML file located in any storage service selected by the user, provided that this storage service implements an API to access the user's files on his behalf such as Dropbox [dro] and many others do. In this way, the service would work with pre-constructed social spheres – contexts – (XML files) transparently to users, but with their permission.

As its second task, *mySocialSphere* would be in charge of providing social

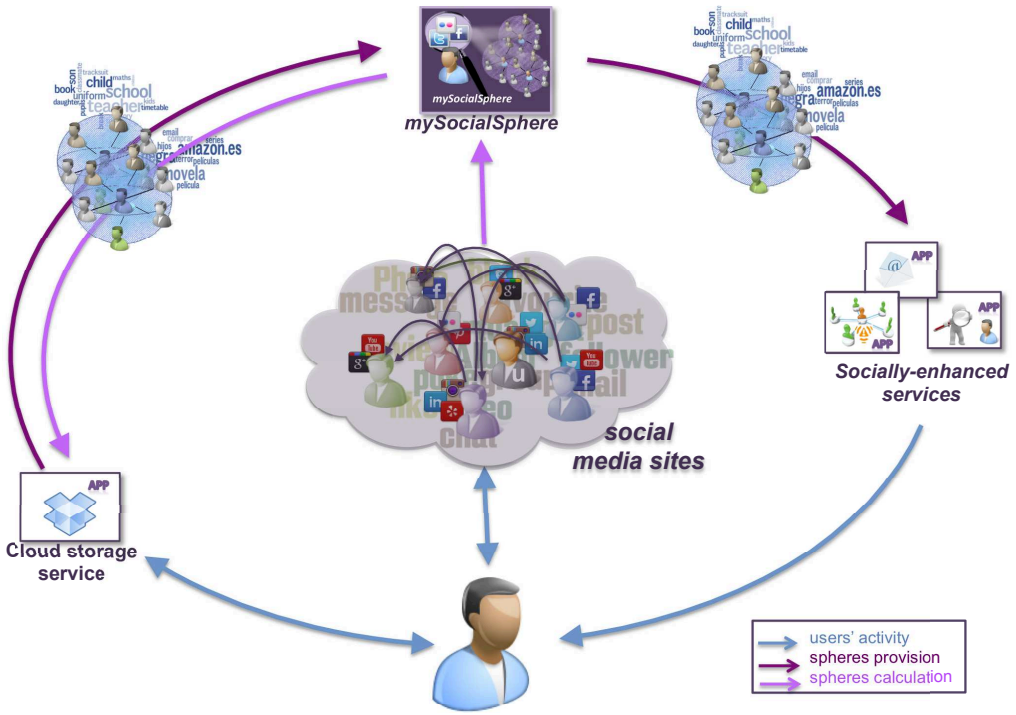


Figure 5.2: *mySocialSphere* Service

spheres to other services as long as they have users' permission. *mySocialSphere*, implemented as a RESTful service, would provide support for other services through an API. Also, to authorise third-party applications – the so-called socially-enhanced services – to access users' social spheres/contexts on their behalf, *mySocialSphere* could use the open protocol OAuth ([Ham10]). OAuth, the protocol behind the most popular social sites (Facebook, Twitter, Google sites, LinkedIn, etc.), provides a method for clients to access server resources on behalf of a resource owner such as a different client or an end-user. It also provides a process for end-users to authorise third-party access to their server resources without sharing their credentials – typically, a username and password pair, using user-agent redirections.

Putting aside the details of the architecture, we have also carried out two pilot experiences to show the suitability of our proposal to develop socially-enhanced services. Specifically, we have used a basic implementation of *mySocialSphere* in both (i) a service in charge of maximising the attention that users receive in social media by letting them know when they should modify the diversity of topics in their conversations and (ii) another service with the capability of finding

trustworthy experts in social media taking into account users' partial views of the whole Web. Apart from detailing how these services work and the experiments conducted to show their efficiency, we end the chapter by discussing other services that might be socially-enhanced using social spheres.

5.3. Application 1: Gaining attention in social media

Social media has turned users into producers of information and, therefore, into competitors for attention. This has also brought individuals the need of managing their online image, applying different methods for generating a distinguished presence on the Web. Although the concept of *online image management* [MMS06] is relatively recent, the *self-presentation* and *impression management* [Gof59] concepts have been studied for decades. These sociological theories state that individuals attempt to influence the perception that others have of them by adapting their behaviour to their audience. In face-to-face situations, the knowledge of our actual audience makes us adapt our behaviour, but in the Social Web, our actual audience is unknown and must be envisioned. In addition, social media allows to reach many more individuals than in traditional face-to-face situations, increasing considerably the potential audience and even making it impossible to determine its real size. In this scenario, the *imagined audience* plays a key role in deciding what content (ideas, news, opinions, ...) is going to be published. So, factors such as self-censorship and adaptation of content to an audience previously envisioned [MB11] influence the writing and posting of messages or any other kind of content in social media.

Given these circumstances, users' publications in social media are both influenced by (i) how they imagine their audience and (ii) their wishes of gaining attention – or being more popular by increasing their audience. Although it may not be possible to alter their perception of their audience, something may be done to help them get attention and gain popularity. To this aim, and using the methodology for extracting social contexts from user generated content explained in Chapter 3, we conducted an experiment *to study the relation between the di-*

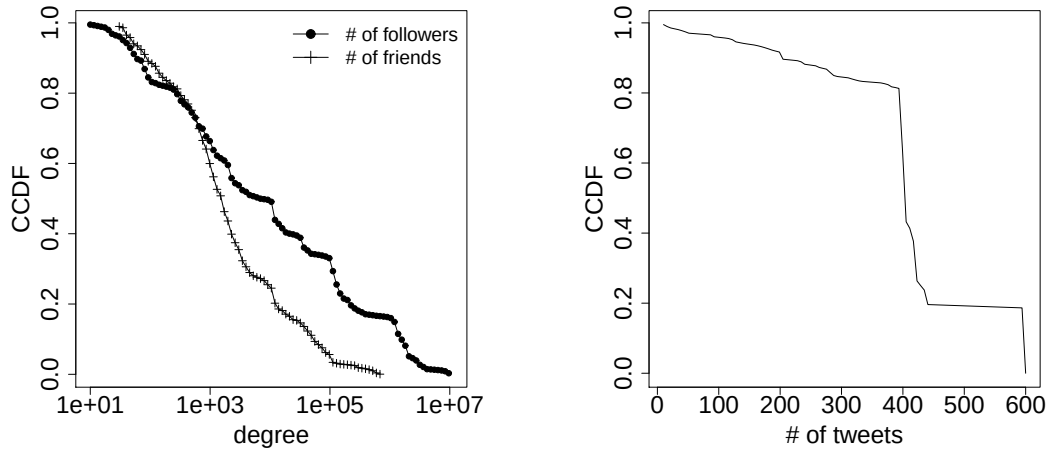
versity of topics talked about by users in their publications and the size of their audience – friends, followers, contacts, etc. That is, to determine whether users deal with few topics when their potential audience is small or, on the contrary, when they have a large audience, and vice versa. In this way, if we could demonstrate the existence of such relation, we might make use of these results in an application that alerts users when they should diversify or specialise the topics in their publication in order to increase their audience. Although more analyses are needed, especially those related to dynamic variations of audience size driven by changes in the topic diversity, preliminary results seem to confirm the viability of such an application.

5.3.1. Experiment

We selected the microblogging service *Twitter* as the social site on which to carry out the experiment. Since the vast majority of Twitter accounts are public and consequently the majority of tweets can be viewed by anyone, the potential audience of a tweet is unlimited. However, its actual audience is composed of only some of the *followers* of the publisher (followed user), being different from the audience that he envisions. In addition, Twitter users do not select their audience, but their audience is who selects them in some discovery process related to the content and topics that they address. For this study, we consider the potential audience of a user u to consist of his followers, i.e. those Twitter users that have established a unidirectional link with him, since u's tweets will be shown on his followers' respective Twitter homepages. Although other Twitter users can also read u's tweets by accessing his profile page or by the Twitter searching tool, we assume that, if user v often visits user u's profile or reads his tweets, v will become u's follower before or after. Below, we detail the experiment and the results achieved.

5.3.2. Dataset

We used the Twitter dataset of Li et al. [LWD+12] obtained by crawling Twitter in May 2011. This dataset contains information about 139,180 users



(a) (b)

Figure 5.3: CCDF of (a) # of friends/followers and (b) # of tweets per user.

including, for each one, at most 600 tweets and his social network (friends and followers). The distribution of potential audiences in this dataset, i.e. the number of users' followers, is similar to the one in Twitter in which up to 7 different orders of magnitude are present. Also, half of the users in the dataset have less than 10^3 followers and 96% have less than 10^4, meaning that most users are *ordinary users* and the number of users with more than 10^6 followers (*celebrities*) is 34. We sampled this dataset to obtain a representative set of users in terms of number of followers. We define six different groups according to the order of magnitude of their audience: users with less than 100, users with more than 100 and less than 10^3 and so on until finally users with more than 10^6 followers, resulting in $2,042$ users in the first group, $76,652$ in the second, $54,571$ in the third, $4,501$ in the forth, 222 in the fifth and 34 in the group of users with more than 10^6 followers. Given that the minimum number of users in a group is 34, we randomly selected 34 users per group, ending with a total of 204 users. Figure 5.3a contains the CCDF of the number of followers/friends per user in the sampled dataset (note that, because of the sampling, the distribution of followers in our sample does not correspond to the one in the whole Twitter [KLPM10]). Figure 5.3b shows the distribution of the number of tweets per user in this final sample, where around 80% of the users have between 400 and 600 tweets.

5.3.3. Particularising the social contexts extraction to this scenario

The mechanism of topic extraction explained in Section 3.2 begins by obtaining relevant words from users' tweets by only considering lexical units that refer to fixed entities with meaning. We used *Stanford CoreNLP* [MSB+14] to filter the text by using POS tagging (which identifies each word part-of-speech category – noun, verb, etc.) and lemmatisation (which identifies each word lemma), only keeping nouns in their citation form. Having obtained users' tag clouds, the *semantic relatedness* between tags necessary for obtaining users' personomy was calculated as *the weighted sum of two different measures*: one based on an external source of background knowledge and the other on the personal knowledge of the user. The former is *Wikipedia Link-based Measure (WLM)* [WM08], the semantic relatedness measure based on the hyperlink structure of Wikipedia. A description of this measure can be found in Section 3.2.3. The latter takes into account the intention of the user when using the term. That is, the intrinsic relation that they acquire for being used together (in the same conversation, same tweet, ...). With the aim of keeping the sense that the user gave to terms, this personal knowledge based measure $sr_u(a, b)$ states that two terms a and b are related for the user u if they appear together in at least, one tweet t of u. Otherwise, there is no relation between them:

$$sr_u(a,b) = \begin{cases} 1 & \text{if } a, b \in t \\ 0 & \text{if } a, b \notin t \end{cases} \tag{5.1}$$

Given the good results obtained with the hierarchical clustering in the closely related study described in Section 3.4, we opted again for the ***UPGMA***, a hierarchical and agglomerative clustering algorithm that yields a dendrogram that can be cut at a chosen height to produce the desired number of clusters (see Section 3.2.3 for a complete description). In order to **select the number of resulting clusters**, i.e. identify individual branches of the cluster tree, we used a tree cutting method that detects clusters in a dendrogram based on its shape: *Dynamic Tree Cut* [LZH08]. This algorithm is based on an iterative process of cluster decomposition and combination that stops when the number of clusters

Table 5.1: Users' topics (clusters) parameters

	average	std. dev.
# of clusters	55.20	25.34
# of representative clusters	28.59	11.71
Silhouette width	0.148	0.03
# of tags per cluster (without r.)	7.07	1.37
# of tags per cluster (with r.)	22.21	27.91

becomes stable. After obtaining a few large clusters by the fixed height branch cut method, the joining heights of each cluster are analysed for a sub-cluster structure. Clusters with this sub-cluster structure are recursively split and, to avoid over-splitting, very small clusters are joined to their neighbouring major clusters. See [LZH08] for a description.

5.3.3.1. Resulting clusters

As a result of applying the topics extraction methodology to the tweets of each user in the sampled dataset, a set of representative clusters of tags emerged, representing topics of interest. However, not all the clusters are representative of a topic, so we considered that, for a cluster to be representative, its Silhouette width [Rou87] must be positive. That is, a cluster c represents a topic when the average dissimilarity (distance) from the point i (member of c) to all other points in c is lower than the lowest average dissimilarity (distance) from point i to all points in any other cluster different from c. The distribution of some parameters of the resulting clusters-topics are shown in Figure 5.4, whereas the average and standard deviation (std. dev.) of these parameters are provided in Table 5.1.

As seen in Table 5.1, the average number of resulting clusters, 55.20, is higher than expected. However, the quality of many of these clusters is not good enough to be considered representative and, after keeping only clusters with positive Silhouette width, their number decreases drastically to 28.59. This methodology still produces a large number of topics per user which, together with the high standard deviation, leads us to clearly appreciate differences in the diversity of topics dealt by some users and others. With respect to the distribution of the number of representative clusters among users (Figure 5.4a), half of users have

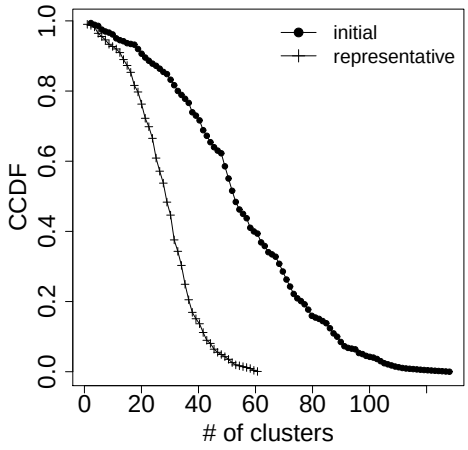
fewer than 30 clusters and users with more than 40 clusters make up less than 15% of the total. In order to prove the significance of our findings, we calculated the Pearson correlation between the number of representative clusters per user and his number of tweets in the sample, finding that they are scarcely correlated (Pearson coefficient = 0.19).

Apart from the number of clusters (topics), the distance between clusters (both intra- and inter- cluster) is relevant to characterise users' topic diversity since the closer the clusters, the less diverse the topics. The average of the Silhouette width (on average for all the clusters of the user) is 0.148 with a standard deviation of 0.03. With respect to the distribution among users, the average Silhouette width ranges from 0 (keep in mind that only clusters with positive Silhouette width are considered) to 0.26. However, the majority of the users (around 80%) have an index between 0.08 and 0.18 (Figure 5.4b), which means that there are not huge differences between users according to the distance among their clusters.

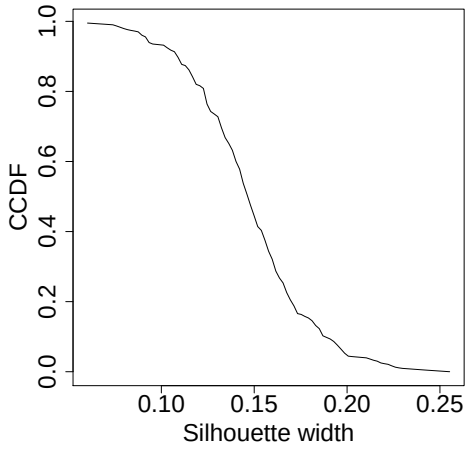
The size of the clusters is also relevant when talking about diversity, since it indicates the relative importance of the topic for the user with respect to the rest of his topics. Figure 5.4c shows that almost all users have clusters of, on average, less than 50 terms and the users with fewer tags per cluster have, on average, 7. But when no repetition of terms is taken into account (*without repetitions* in the figure), the differences among users with respect to the average of terms per cluster are drastically reduced, since almost all users have clusters with, in average, between 5 and 10 different terms. This is also observed in Table 5.1 and specifically in the difference between the standard deviations when taking into account tag repetition in clusters (*with r.*) and not (*without r.*).

5.3.4. Relating content diversity with audience size

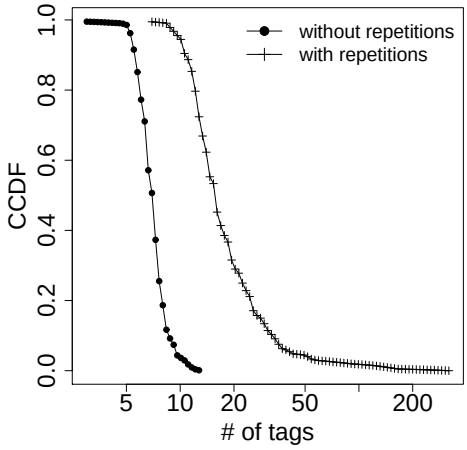
We define the potential audience of a user as the set of Twitter users that follow him and his topics as the clusters of terms resulting from applying the methodology explained in Section 5.3.3. We calculated the Pearson correlation coefficient between the number of representative clusters (topics) and number of followers for all the users in our dataset. As the number of followers involves



(a)



(b)



(c)

Figure 5.4: CCDF of (a) # of clusters, (b) Silhouette width and (c) # of tags per cluster for all users in the dataset.

different orders of magnitude, we calculated this correlation between the number of clusters and the logarithm of the number of followers, obtaining a value of 0.218. This positive correlation means that users with many followers tend to have higher topic diversity than users with less followers. But, the increment of topic diversity is not fixed with the increment of the number of followers.

In order to put aside the influence of the accurate values of number of followers, we opt for an analysis using groups. We group users according to the order of magnitude of their number of followers, starting at 100 followers. Figure 5.5a shows the boxplots of the number of topics dealt with by the users in each group. In view of the results, users with more than 10^6 followers (*celebrities*) clearly have a higher diversity of topics than the rest of users in the dataset (34.06 on average, versus 31.44 in the case of users with between 10^4 and 10^5 followers or even less than 30 topics for users with less than 10^4 followers).

Although grouping users according to the order of magnitude of their number of followers seems a suitable classification, from Twitter's view classifying users into *celebrities* and *ordinary users* makes much more sense. As the limit for being considered a celebrity according to followers is not clear, we did a new classification of users into three different groups: users with less than 10^4 (*ordinary users*), users with between 10^4 and 10^6 and users with more than 10^6 followers (*celebrities*). Results in Figure 5.5b show that, in average, the number of clusters (topic diversity) is different for the different groups, being the lowest in the case of users with less than 10^4 followers (26.08) and the highest in the case of users with more than 10^6 followers (34.06). With all of this, what is clear is that the larger the audience, the higher the topic diversity.

Finally, the boxplots of the distance between clusters – Silhouette width – and the number of terms – tags – per cluster, are provided in Figures 5.5c and 5.5d respectively. Although the average Silhouette width is similar for the users in the different groups (around 0.15), the variance is higher in the case of *ordinary users* than in the case of *celebrities*, being 1.26×10^{-3} for *ordinary users* and 6.85×10^{-4} for *celebrities*. With respect to the number of tags per cluster, Figure 5.5d shows that the average number of tags per cluster is similar for all the users, but the variance is higher in the case of *ordinary users* than when the users are *celebrities*. This is in consonance with the results obtained in terms

of number of clusters since, when considering approximately the same number of tweets per user, *ordinary users* tend to talk a lot about fewer topics, while *celebrities* talk about more diverse topics with lower intensity.

5.3.5. Final remarks

The experiment showed that *ordinary users* talk about fewer topics, but with more intensity, than *celebrities*. This confirms that users' behaviour is affected by their audience as expected given the theories of *self-presentation* and *impression management* [Gof59]. It could seem that users with a large number of followers (*celebrities*) tend to minimise their content diversity, dealing only with those topics that have made them famous (sportsmen about their own sport, politicians about their own party, actors about their movies, etc.). However, as Marwick and Boyd claimed [MB11] other factors come into play when *celebrities* post tweets. Apart from tweeting about their own interests and likes, *celebrities* make efforts to discover the interests of their followers to tweet accordingly and also to satisfy their sponsors promoting their products by sponsored tweets. They are unconsciously forced to keep the balance between keeping their authenticity, keeping their followers and keeping their sponsors' support, which would explain the inevitable increment of the content diversity of their tweets with respect to *ordinary users*.

5.4. Application 2: Finding trustworthy experts in social media

The application proposed in the previous section helps users gain attention in social media by presenting relevant and useful content to their contacts – followers. In such an application, contacts are considered passive subjects, restricting their consumption to only those pieces that people they follow choose for them. However, these contacts often need to find information or advice that goes beyond this previously selected content. In the real world people typically ask their friends, colleges, relatives, etc. for advice when planning to purchase a new item.

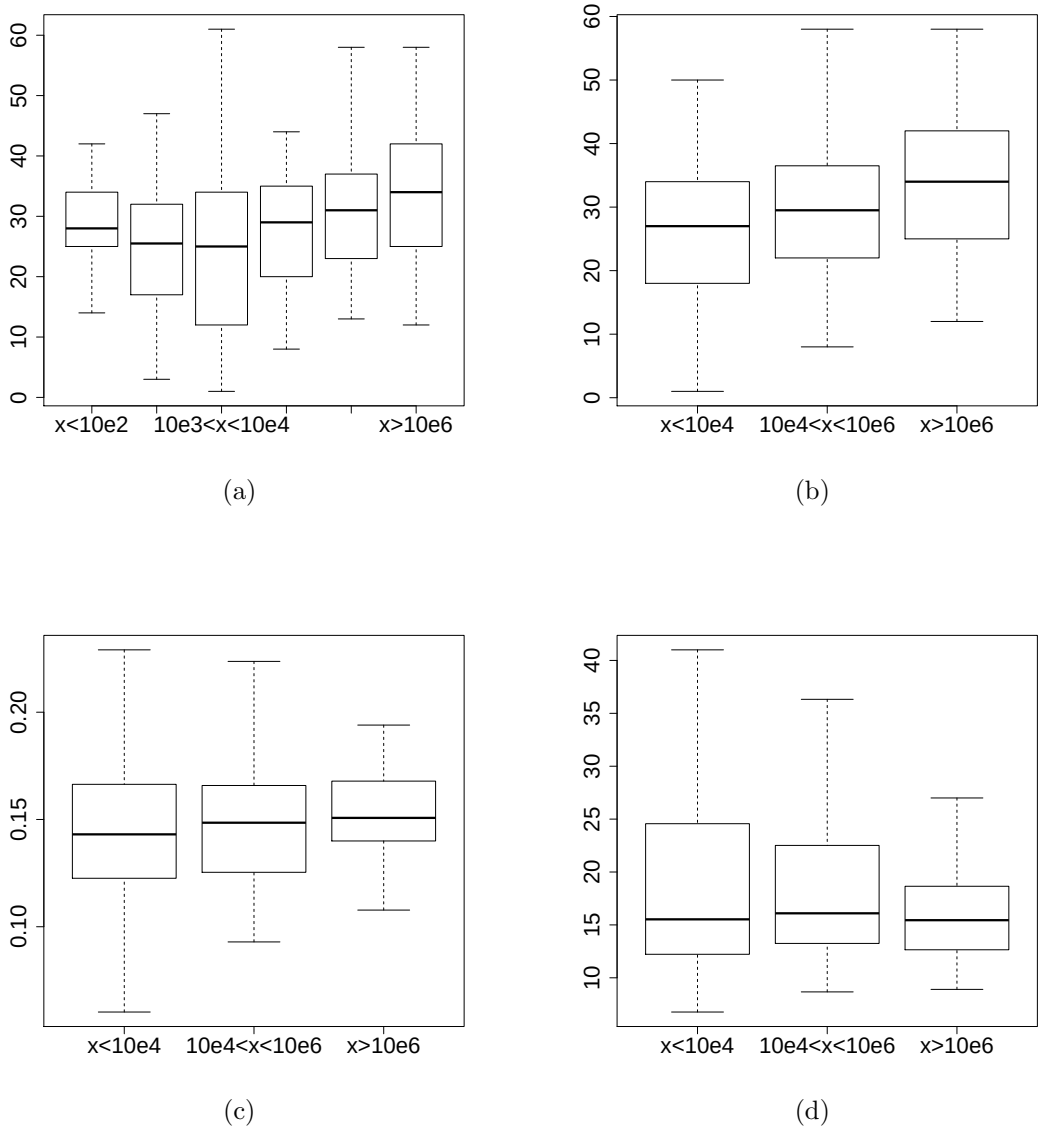


Figure 5.5: CCDF of (a) # number of clusters per group for users classified into 6 groups; (b) # number of clusters per group, (c) Silhouette width and (d) # of tags per cluster for users classified into 3 groups.

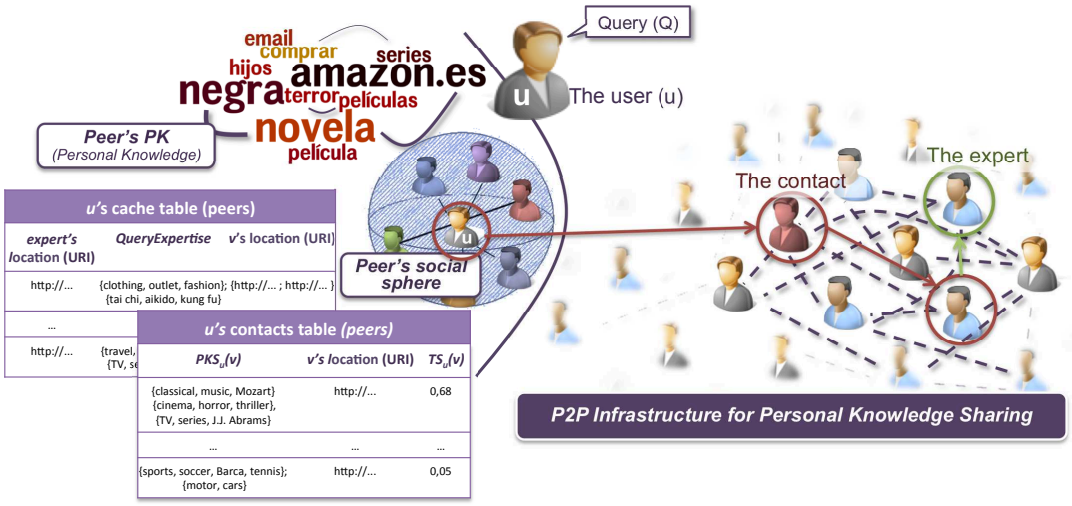


Figure 5.6: The scenario for finding experts in the social P2P network

Translating this into the online world, and given that our social spheres model contains information about the contacts and interests – or knowledge – of the users, it seems suitable to take advantage of these social spheres to assist users to find – as opposed to consume – trustworthy and useful information.

In this online scenario, it is common for individuals not to have any contact expert on the topic – item – on which they are seeking advice. However, these contacts may have some other contact that could advise the former. This is the basis of the second socially-enhanced service that we propose in this chapter: an application that takes advantage of users' social spheres – contexts – to find trustworthy experts in social media that can provide them with useful information. Such an application is based on a social P2P protocol that allows individuals to reach the knowledge of other peers by following the ties or links in their social spheres.

5.4.1. Application overview

Our social P2P approach for knowledge dissemination, retrieving, etc. relies on social spheres and social contexts of individuals to establish the peers architecture. According to the scenario we introduce in Figure 5.6, the user u owns his *social sphere*, which models his online life – his contacts and strength of relations

with them – and *social contexts*, that model his interests or knowledge – topics in which he is considered expert for our application – manifested in social media. This information, spheres and contexts, is local to the user and is only allowed to be shared with other users inside his spheres – his social contacts. Whenever he needs to access to knowledge outside his sphere, he takes advantage of these contacts. That is, we see contacts as social peers and, as other social P2P approaches, we build an overlay network based on the peers' social characteristics, respecting their privacy.

Under these premises, the user u firstly tries to find an *expert* on a specific matter, and so sends the query message to his social peers, which is automatically routed on the overlay network until the expert is found. We define an expert on a given topic as someone who talks about this matter and, consequently, is interested in it. Therefore, an expert is a key person because he probably knows the answer to questions related to the topics in his knowledge or because he knows someone who can give them to us. After that, u sends the question to the expert. Although it may be possible that there is direct contact between them, we provide a mechanism based on a chain of intermediate peers who participate in propagating the questions to the expert, with the aim of guaranteeing that all the communications take place between two peers that share a link in at least one social network site, thus ensuring the communication is possible. In order to reach this scenario, our proposal takes advantage of the social spheres introduced in this dissertation to design a routing algorithm (Section 5.4.2) to locate experts and supplementary mechanisms to articulate the return information about the expert and the posterior formulation of the query (Section 5.4.3).

5.4.1.1. Users' local knowledge

Users have a biased and egocentric vision of the online world – the Web, and consequently of the social network structure. As indicated, this perspective is focused on (i) the colleagues with whom the user interacts (ties that form his social sphere) and (ii) the subjects about which they talk (social contexts) and in which they are supposed to be experts. With this information the knowledge about each of his colleagues v that the user locally maintains is composed by

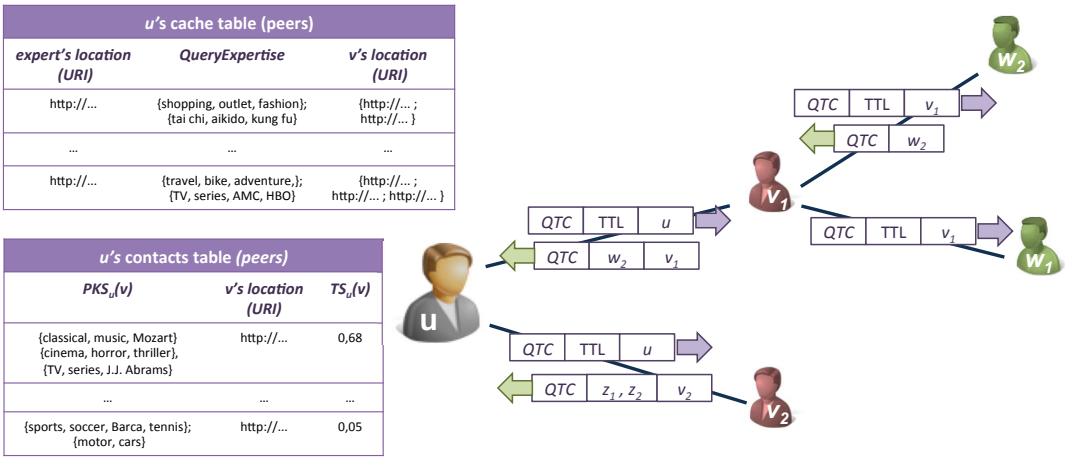


Figure 5.7: u's information for the distributed search

a set of 3-tuples: $\{v, TS_u(v), PKS_u(v)\}$, where $TS_u(v)$ represents tie strength (Section 4.1) and $PKS_u(v)$ is v's personal knowledge summary, i.e. the subjects or themes on which v shows a certain kind of expertise. u knows the topics on which v is an expert from the content of their interactions on social media sites, i.e., from the set of u's social contexts in which v is involved, $c_{u|v}$ (Section 3.1). With the aim of efficiently storing information about the peers to route a query, we propose that u stores only a brief summary of the knowledge of his peers v. So, $PKS_u(v)$ is obtained by considering the three tags with the highest multiplicity in $c_{u|v}$.

5.4.2. The search algorithm

The search starts by user u asking a specific query in natural language, for instance, "What is the dog vaccination schedule?", which is processed by using a natural language processor such as *Stanford CoreNLP* [MSB+14] to extract a set of relevant tags: a query tag cloud, denoted QTC, that constitutes the starting point of the search. Since we have defined a user-centred knowledge model, each user maintains his own information coming from two different sources (Figure 5.7): (i) his own social sphere and social contexts, and (ii) a cache table keeping information about experts out of his sphere (second table in Figure 5.7), as a consequence of the P2P search algorithm. u's contacts table, stores the description, location information and tie strength of each user v to whom u is

related in any social media site. u's cache table stores information about each expert that u has found as result of the execution of the P2P search algorithm: his location information, the tag cloud of the queries on which he is supposed to be expert – $QueryExpertise$, and a list of the colleagues in u's social sphere through whom the expert was located.

Once the query is formulated and processed, the search algorithm tries to find at least one expert among the peers user u knows (tables in Figure 5.7). If this attempt fails, the query is forwarded to other peers to find this information out of u's scope. Finally, whenever a user v has a set of experts for the target query, he returns this information to the user u, so he can address the question directly or indirectly to one of them.

Peer ranking. In order to find those peers in u's peer set (social knowledge and cache table) whose expertise is close to the target matter, the algorithm firstly computes a peer ranking: (i) by comparing the query tag cloud (QTC) to the personal knowledge summary ($PKS_u(v)$) of each colleague v in u's social knowledge and (ii) by comparing the QTC to the $QueryExpertise$ of each expert in the cache table. For the comparison, we propose a similarity measure between two elements e_i and e_j (which might be any tag cloud previously described: $PKS_u(v)$, QTC or $QueryExpertise$). This comparison not only takes into account direct tag matching but also relations between tags in a folksonomy. This folksonomy-based similarity $FolkSim(e_i, e_j)$ takes into account those terms that, despite not being included in both tag clouds, are related in the folksonomy:

$$FolkSim(e_i, e_j) = \frac{\sum_{t_k \in TC(e_i)} w(t_k, e_i).max\{w(t_l, e_j)).r_{kl} | \forall t_l \in TC(e_l)\}}{\sqrt{\sum_{t_k \in TC(e_i)} w^2(t_k, e_i)} \sqrt{\sum_{t_l \in TC(e_j)} max^2\{w(t_l, e_j).r_{kl} | \forall t_l \in TC(e_l)\}}}. \tag{5.2}$$

This folksonomy-based similarity does not average the weights of the same tags in both tag clouds, but the more relevant paths in the folksonomy between the tag t_k in the tag cloud of e_i and the tag t_l in the tag cloud of e_j. Therefore, we select the maximal value of $\{w(t_k, e).r_{kl}\}$, with r_{kl} being the relationship of t_k

and t_l in the folksonomy. We also use a semantic relatedness measure between tags to obtain the relations in the folksonomy.

Forwarding the query. Having the peers ranked according to QTC, the algorithm stops if there are users whose comparison results are higher than an established threshold Th_{pks}. Otherwise, u forwards the query (together with the time-to-live, TTL, and the u's URI) to a subset of his peers (see Figure 5.7), which is selected according to the following 3 criteria:

1. All those peers having comparison values higher than $Th_{approx_{pks}}$. Thus, they are not experts on the query, but their knowledge is close enough to that required. So, they constitute a target group of peers who talk about topics related to the query.

2. Those peers in u's social sphere having the highest tie strength indexes, under the premise that the more active a user is, the more possibilities of finding an expert among his peers.

3. A set of randomly selected peers in u's social sphere having low tie strength indexes, with the aim of broadening the search and increasing the possibility of finding experts out of u's social range of action.

So, the whole procedure is as follows. Whenever a peer v receives the query message he first checks if himself is an expert and, if so, the search ends (like w_2 in Figure 5.7). Otherwise, v checks the degree of expertise of his peers (table cache and social sphere) on the query's subject by calculating the similarity between $PKS_u(v)$ ($QueryExpertise$) and the QTC, using Equation 5.2. If this comparison results in a set of peers having a higher value than the established threshold Th_{pks} then the search ends because a list of experts has been found (like v_2 in Figure 5.7). If the comparison does not succeed, v forwards the query according to the previous criteria replacing in the query message u's URI by his own identification (URI) until the query TTL expires (like v_1 in Figure 5.7).

5.4.3. After finding the expert

Having explained the routing algorithm, it remains to describe mechanisms to (i) propagate back experts' locations, (ii) update information stored by the peers and (ii) send the question to the selected experts by using publishing mechanisms in social media sites.

Answer propagation throughout the peers network. Once a peer v has a list of potential experts to return (himself or a subset of peers in his social sphere or his table cache), the answer message is created, including the following information: the QTC, the set of URIs of the potential experts, and v's URI (like v_2 in Figure 5.7). This message is propagated throughout the peers network following the same path that the query has taken, but the other way round. The answer message is always sent to the colleague who forwarded the query to v, who replaces v's URI by his own URI (like v_1 in Figure 5.7) and proceeds in the same way starting a chain that ends with the user who forwarded the query in the first place: user u. Therefore, u has the list of potential experts and now he is ready to ask the question.

Updating both social spheres and table caches. Information in cache tables is updated each time a user v receives a list of potential experts from any colleague in his social sphere: he checks one by one if the experts are already in his cache table or not and updates the data, i.e. the query on which the peer is supposed to be expert, his location information and the colleague who sent the information. Finally, and with the aim of maintaining a reasonable cache table size, we introduce a forgetting mechanism to remove those entries that have not been used for a long time. Data in social spheres is also updated using the answer propagation mechanism: whenever a user v sends a response message to one of his colleagues, he also sends his updated personal knowledge summary, a summary of his social contexts.

Asking the question to the experts. The whole procedure ends by user u asking the query, or any other question about a related issue, to one or more of the experts he has located. There are two main issues here: (i) how u selects the most appropriate expert to ask the question and (ii) how the question is sent to the expert. Regarding the former, having a list of experts to ask the

questions, user u has the opportunity of selecting a specific subset. Although in the current solution, the question is sent to all of them, we are working on incorporating some mechanism of reputation management that can aid the user in selecting the expert. Regarding formulating the question, there are two different ways to proceed. User u might contact the expert directly by using one of the mechanisms supported by a social network site or by using other communication strategies (e-mail, phone, etc.). Although this solution is efficient, it may not be the most suitable: (i) because u is trying to contact a total stranger who could be reluctant to give him an answer and (ii) because some social media sites do not allow direct communication with users out of our social circles. The other option is propagating the specific queries or questions throughout the overlay peers network using the same peers-path used to locate the expert, but in the other way round. This is the reason why we propose to store the information about the colleagues who have forwarded the query message. Using this solution, we avoid the two aforementioned problems of communication in the social media site and/or cooperation from the expert.

5.4.4. Pilot experience and final remarks

In order to validate our proposal, we deployed a Facebook application that improves the functionality of *Facebook Questions* [Facd] letting users get recommendations from his friends and other people. Our application *SQ* (*Smart Questions*) allows users to make a question and automatically routes it according to the relevant tags in the question and the personal knowledge and tie strength of social peers according to the routing scheme previously described. With this aim, the application asks users for their permission to use the service *mySocialSphere* on their behalf, and posts the question to the wall of the social peers selected by the routing algorithm. We recruited (under)graduate students from the University of Vigo to use the application and give feedback about the perceived utility of the responses and their willingness to rely on a response to take some decision.

We compare the students' scores in *Facebook Questions* with those in our application case. The perceived utility of the responses doubles in the case of *SQ*, although the number of responses is approximately 1/3 less in the *SQ* case.

Unfortunately, the second parameter (willingness to rely on the response) worsens when the expert is outside the social sphere of the user. In the context of Facebook in which the experiments were performed, we can say that students were reluctant to take decisions according to the expert's response when the expert was neither a "friend" nor a "friend of a friend". This observation demonstrates the need to incorporate some form of reputation management in our proposal. In addition, students expressed significative disappointment in the fact that some selected experts took days to give a response on the Facebook wall. In some cases, the response becomes less useful with the passage of time.

In conclusion, our social P2P approach to find experts is based on the idea that people build social relationships with each other and these relationships may help people find appropriate information or services more effectively. Using the peers' knowledge and social spheres we can define a dynamic overlay networks adapted on the fly to the requested matter. Furthermore, because of the random factor in peer selection, other problems like deadlocks or endogamy issues are avoided.

5.5. Discussion

This chapter helps to fill the gap between the theoretical explanation of our model of social spheres and a feasible practical implementation of such an intermediary online service; as well as showing the benefits of using such a service in socially-enhanced technological social applications. The services for helping users gain attention and find trustworthy experts in social media are only some of the services that can be socially-enhanced using our social spheres model. Below, we discuss other services or applications that have been, or might be, personalised using these spheres. In order to organise the presentation of these services, we classify them according to the data flows that can be identified in the Web: (i) the access to content (consumption), (ii) the production of content (sharing) and (ii) the organisation of contacts in the Web (contact management).

5.5.1. Consuming resources

The massive availability of services in the Web poses challenges and opportunities comparable to those of the massive availability of information on the Internet. To this aim, software solutions arise to assist users in finding services to satisfy their specific needs, interests, etc. Some of these are, for instance, those related to recommendation, social marketing or attention management.

Recommendation. Collaborative filtering, as the most successful approach to recommender systems ([Bur02, MLdlR, AT05]), is based on the premise that users who have historically had similar interests will probably continue having them into the present. An important issue in these systems is finding a set of users, known as a neighbourhood, that have a history of agreeing with the target user (having rated services similarly, tending to use similar services, etc.). Several authors, like O'Donovan and Smyth [OS05], have improved neighbourhood formation by taking into account, apart from similarity between profiles, social influence or trust between users. As in real life, when we look for a piece of advice (on health, commerce, learning, etc.) we often turn to our friends, on the basis of our implicit trust in them. In the case of collaborative filtering systems, knowing the tie strength of the neighbourhood would improve the effectiveness of the recommendation. Furthermore, the contextualised social spheres could be useful when the scope of the items to recommend is similar to some user's life context.

In the same line, Blanco et al. [BLPM12] propose a system that promotes free tourist resources and activities through the most influential users in Facebook and Twitter. The majority of influence metrics provide values that take into account the user's contacts' influence, considering all of them to have the same importance. They propose to improve previous metrics on the basis that influence depends on users' real contacts (those with strong ties). With this enhanced application they found that the number of free activities the users accepted was somewhat higher than using directly the influence values provided by online tools. Also, in [FDS14], we propose to consider tie strength between families to improve a parental monitoring system over DVB-IPTV. Specifically, we propose a collaborative filtering system that infers the IPTV content that should be blocked

from tagging and blocking data provided by other parents. As the decision about blocking comes from filtering of other parents' opinion, the social spheres have been used to improve the neighbourhood formation taking into account, apart from similarity between profiles, tie strength between parents.

Group recommenders aim to suggest interesting items to a group of users instead of an individual. For instance, [GXL+10] propose a solution which firstly generates recommendations for each member of the group and later merges the individual recommendations, selecting those that are the most suitable for the whole group. For the merging process, it uses a consensus function which depends on, apart from other information, the strength of the ties between the users in the group. As in the traditional case, group recommenders might also be benefited if they took into account the tie strength between users in the context of the items to be recommended that *mySocialSphere* provides.

Social marketing. The integration of web publicity and social media is emerging as a new trend in marketing. For instance the authors of [ZJC11] found that brands' engagement in communication on Twitter enhances consumers' engagement in word of mouth communication. However social media is not enough on its own and being able to identify and target the most influential users is essential to improve effectiveness. Selecting these key users in a wide graph is an interesting task that has received a great deal of attention in recent years. Many of the algorithms to calculate users' influence in social networks are based on the premise *the more influential your friends (contacts) are, the more influential you are*. However, a friendship on Facebook does not necessarily imply the existence of a relationship or, at least, the existence of a strong tie which can guarantee influence to some extent. Consequently, a user's influence may not affect the influence of his contacts if they are not related.

Prominent examples of social marketing are *Facebook Ads* [Facb] and *Groupon* [BMPZ11]. *Facebook Ads*, for instance, uses Facebook as its advertising medium, allowing the manual creation of advertising campaigns and selection of target users, filtered according to data in their profiles. Ads in the campaign will be linked to the Facebook homepages of these target users. However, to the best of our knowledge we believe that this tool is not aware of users' activity on Facebook that, undoubtedly, would improve the effectiveness of the product. That

is, *Facebook Ads* could be enhanced by selecting, as receivers of the campaign, those users with many strong ties in their social spheres in the context of the campaign (defined by similar tags to those used by the trader to describe the advertisement). In accordance with our model, a user has a social sphere in a specific context if he has relevant social activity in this context. Besides, as our social spheres are created from activity data in more social sites than Facebook alone, the selection of receivers of the campaign would even be more effective, as we demonstrated in Chapter 3. On the other hand, *Groupon* is a deal-of-the-day website which offers discount coupons usable at several companies. Its business model lies in that if a certain number of people sign up for the offer, then the deal becomes available to all; but if the predetermined minimum is not met, no one gets the deal that day. In order to increase its customers, *Groupon* allows users to refer friends to the site and, in return for friends buying their first coupon, get credits to spend in future coupons purchases. In this case, *mySocialSphere* may help users to propagate coupons among their contacts. To this aim, it would infer, from the products in the coupon, those interaction contexts which best fit, making the coupons most interesting to users. In this way, coupon redemption would increase and, thereby, also the users' credit.

Attention management. Some e-mail readers, such as *Mail* of *Mac OS X*, allow users to define smart mailboxes, sorting mail into different folders depending on their content, header, sender, etc. However, defining and updating mailboxes are tedious tasks. But as many received messages are from users' contacts on social sites, knowing the strength of their ties would help to suitably define smart mailboxes parameters and, even, to prioritise incoming messages. So, an application to manage users' mail might benefit from social spheres. This application would suggest (and even create) smart mailboxes which include messages from contacts in a specific social sphere. Also, if the senders are not the addressee's contacts on social sites, the application would allow users with strong ties to share their smart mailbox parameters (and, therefore, the senders' identity), turning it into a distributed and collaborative application. In this line, the application might also work as an application of collaborative spam email mitigation along the lines of [GKF+06]. That is, if user u marks as spam messages from user w, and user v has a strong tie with u, v's e-mail reader will mark as spam all messages received from w.

Similarly to e-mail readers, social media dashboards, such as for instance HootSuite [Hoo] and TweetDeck [Twe], are examples of tools in charge of managing users' attention. Social media dashboards are platforms used to organise users' updates in different social networks. They allow users to inspect their contacts' activity and post new content without connecting to social network sites. However, we believe that social media dashboards do not exploit their full potential. For instance, if they were aware of users' social environment, they could filter contacts' activity. They would show only updates from contacts with strong ties and even organised by context. That is, they would do something similar to *Facebook News Feeds*, but for every social site and using social spheres to show to users only updates about important issues in their social relationships.

Attention management is also a key issue for Ambient Intelligence. For instance, in [SKQ09] Shannon et al. found that users' interactions in social networks have a recurring rhythm, i.e. users tend to interact with their contacts with a certain rate and regularity. With this finding, they propose an application that monitors users' rhythm of interactions (phone calls and SMS text messages) to occasionally recommend them to contact certain friends in order to keep their social network in a healthy state. The increasing number of Social Web users suggests the need for a similar application that can be easily deployed using social spheres.

5.5.2. Sharing resources

In the Social Web, users post content related to different topics or interests and share it with a subset of their contacts (usually related to them in the context of the content). However, specifying this subset of users is a time-consuming activity, which could be alleviated using our social spheres in the context of the content. These socially-enhanced sharing services could be, for instance, selective posting in social sites services, P2P systems and other file-sharing systems.

Selective posting. Using the functionality offered by *mySocialSphere*, social media dashboards could, apart from filtering users' activity, suggest them lists to share new content with. It would be similar to a proxy between users and, for instance, Facebook: users would write the content to share in the platform (photo, wall-post, etc.) with its associated information (photo title, people to

tag, text of the content, description, etc.) and, depending on this information, the dashboard would suggest one or several lists with which share the content. For instance, when one user wants to upload photos of the friendly basketball match he played last weekend, the application would suggest that he share them with his contacts in the *FriendsLists* of users with strong ties in "basketball", "sports" and "friends" contexts. In addition, many web sites include buttons to share content in online social sites, such as Twitter Buttons [Twib] and Facebook Buttons [Facc]. Along the same lines as the previous proposal, a provider in the Web could offer a social button to be incorporated into web sites. This social button would (i) analyse web site content and (ii), according to social spheres, decide in which social network to post the content and with which contacts list (or lists) to share it.

P2P and other file-sharing systems. Knowing social spheres might also be useful to share files or folders in storage services (for example, *Dropbox* [dro]). Our social spheres may be used to find suitable users with whom to share files uploaded to Dropbox accounts. Also, in the case of P2P systems, when a peer is looking for other peers in the system, knowing who the peers in his social sphere are could speed up the search procedure. Moreover, one user may be interested in getting files from specific contexts (for instance, photos about the last world basketball championship). In this situation, discovery would be improved by looking for the file among the peers in a social sphere associated with that context. If the file has not been found, the query could be propagated through the contacts in these contextualised spheres.

5.5.3. Contacts management

Interaction networks are made of links or ties between users who regularly interact through online social sites. However, all these social ties are not considered equal from the user's view and, often, he needs to have them organised. In other online applications, such as e-mail readers, users also need to have their contacts organised according to some criteria. Consequently, some tools help users in this organising task such as, for instance, user lists on Facebook and Twitter.

Facebook allows users to create lists of contacts (*FriendsLists*) with whom

to share specific content. Creating *FriendsLists* entails the user assessing their contacts to include them in one *FriendsList* or another. Moreover, Facebook relationships change over time and, consequently, the contacts to include in the lists could change too. For this reason, Facebook automatically creates smart lists whose members are filtered according to profile similarity. In this respect, combining similarity with tie strength may be useful to improve these smart lists, so that they are composed of contacts who share profile characteristics and, at the same time, are usually connected (even in other social sites). Consequently, our social spheres would be useful to applications in charge of suggesting, and creating various *FriendsLists* depending on their interactions and the different contexts in which interactions happen (close friends list, workmates list, acquaintances list, friends of hunting club, etc.).

The same idea may be applied to Twitter where users can also group their followees in lists. Twitter lists, unlike the Facebook ones, are not related to levels of privacy, but they are guided, mainly, by topics used by *twitterers* or by the context in which followers and followees are related. As in the Facebook case, creating and updating lists are tedious tasks. For that reason, our social spheres could be used to classify users' followees (or suggest possible classifications) into different lists depending on the contexts in which users relate and their tie strength in the considered context.

5.6. Summary

In support of the thesis that an intermediary model of the user constructed by properly mining user generated content in social media can be exploited to create or improve technological social applications, in this chapter we have described a tentative practical implementation of such a model into a software-as-a-service application, *mySocialSphere*, that provides social spheres on request. While the actual implementation of a service like this faces privacy issues that are beyond the scope of our research, it is clear that it is necessary to protect users' data. We believe that solutions for this protection should guarantee that users are the only owners of their data and free to store them wherever they decide, which does not prevent them from sharing their information with whatever entity (service or

user) they wish.

We have also detailed two applications that might be socially-enhanced using this service and experiments or preliminary results towards their full development. These applications, devoted to helping users gain attention and finding trustworthy experts in social media, are not the only ones to be socially-enhanced using the proposed model, but a wide range of others would benefit from our social spheres service. With this, it seems clear that is feasible to implement a service that takes advantage of social media sites to socially-enhance or personalise external services and that there are a variety of services that might be interested in externalising their provision of personalisation by using an intermediary service like that proposed in this chapter.

6

Conclusions and further work

6.1. Thesis summary and contributions

In previous chapters we have presented and discussed the results of our research in relation to the thesis that *an intermediary model of the user constructed by properly mining user generated content in social media can be exploited to create or improve technological social applications*. These results, as well as the contributions they represent to build on the existing research outlined in Chapter 2, are summarised below.

In Chapter 3, we proposed a **methodology to extract users' interests from textual publications that individuals freely post on social media sites**, which has the potential to be developed without any a priori knowledge about the number and categories of interests, nor a priori knowledge about the users for whom the extraction is applied. We then evaluated our methodology, based on data mining and natural language processing techniques, by means of

(i) a user study on Facebook and (ii) using Twitter hashtags. Our findings from the former study showed that the consideration of the content spontaneously generated by users allows the accurate prediction of user-preferred deals, demonstrating that our methodology of interest extraction could be applied to improve social publicity strategies. But this study has also brought insights into how users are not reliable when they are asked about their interests. The latter, the study on Twitter that overcame the limitations of unreliability of participants and lack of data of the former, revealed that (i) a hierarchical clustering (UPGMA) is the best performing algorithm and (ii) the suitability of using Silhouette width to select the input parameters to the clustering algorithms.

This research into the extraction and application of users' interests from user generated content in social media with personalisation purposes builds on earlier work that considered users' content, mainly from users' profiles, to personalise applications created on the same site from which the profile was extracted. But our contribution goes a step further and, instead of focusing on structured content from users' profiles, it is based on successfully mining the unstructured content that users freely post on the given social media sites. Also, instead of imposing the constraints of a previously fixed classification of interests, our solution is based on a bag of words representation. This, apart from involving a high granularity in the definition of interests, means that our social contexts can be easily integrated into almost every service that wishes to be personalised or socially-enhanced.

In Chapter 4, we presented a **user-centred measure of tie strength between two individuals from evidence of interactions in social media**. We described this methodology to assess the closeness that one user perceives of his relationship with others using evidence of their interaction activity on different online social sites (the tie strength between two users, u and v, from u's perspective). We then showed the validation of our measure by means of a user study on Facebook where participants were asked to classify their contacts into groups of closeness, and how our measure was used to predict the group into which each contact was classified. Our findings demonstrated that this measure produces an acceptable classification of individuals' contacts into circles of closeness and that users are not as reliable as expected when they are asked about their Facebook relationships.

This work on measuring the strength of the relationship between two individuals from the perspective of one of them builds on previous work focused on assessing what Granovetter [Gra73] called *tie strength* between two individuals. Initial studies on measuring tie strength were mainly conducted through surveys of human participants, characterised by providing only a limited and very static view. But now the emergence of social media and its widespread use make data much more available and such research feasible to conduct. Contrary to previous works focused on analysing data from only one site [WBS+09, BRMA12, ZWF+12], we consider the complete view of users' online lives by taking into account their evidence of interactions in all the sites where they have created an account. Also, to accurately assess the closeness between two individuals, we consider a broad range of factors such as distinct types of interactions and contexts, the time in which interactions occur, the people involved in them and the frequency of the interactions with the rest of the user's contacts.

In Chapter 5, we described **how our social spheres model could be easily implemented as an online service following the software-as-a-service paradigm**, a service that would be in charge of building, managing and delivering the spheres on request and always with users' permission. Implementing the social spheres model into a service that follows the software-as-a-service paradigm and provides support for other services through a REST API may increase applications' willingness to consider it as an external tool for providing their users with satisfactory personalised experiences and without detriment to privacy. We then analysed the possibility of using the spheres in two socially-enhanced applications in charge of helping users (i) gain attention and (ii) find trustworthy experts in social media. For the former, we detailed our experiment to study the relation between the diversity of topics discussed by users in their publications and the size of their audience. The discovery of such a relation demonstrated the suitability of using social spheres – social contexts – in an application that alerts users when they should diversify or specialise the topics in their publication in order to increase their audience. For the latter, in order to overcome users' isolation caused by their limited view of the system – they are only aware of what happens in their social spheres – we detailed our social P2P proposal that takes advantage of social spheres – both in terms of tie strength and contexts – to help people find appropriate information or services more effectively. As these are

not the only services to socially-enhance, we ended the chapter reviewing other applications that might be personalised using social spheres.

The detailed description of the social spheres-based service as well as the relation of applications to socially-enhance with it constitute the final proof of the feasibility of using user generated content in different social media with personalisation purposes. Also, the externalisation of the personalisation provision by taking advantage of the service-oriented computing paradigm involves a considerably reduction of services' workload, allowing these services or applications to focus only on their main task. Finally, users' sensitive data are owned by users who are willing to share them only with services they trust, which comes to the aid of the increasing concerns about privacy in social media.

In summary, our research provides evidence that it is feasible to construct an intermediary model of the user, both in terms of his interests and his social life, by properly mining user generated content in social media, and that this model can be useful to personalise or socially-enhance a broad range of technological social applications, either existing or to be developed, from recommender systems to e-mail readers.

6.2. Directions for future research

The research covered in this dissertation builds upon the foundation of previous work, but also raises some open issues to be progressively tackled in future research in the area.

Social media users talk about topics that they find interesting during all their lives – permanent interests, but also about others that they only find interesting at some moment because they are trendy, novel, etc. – temporary interests. **Discerning between permanent and temporary interests** would surely be highly appreciated by many online services. For instance, by recommender systems in order not to offer old recommendations that could bother people. In Chapter 3 we explored the possibility of modelling users' manifested interests in the online world by means of bags of representative words; although the impor-

tance of users' interests – contexts – may be determined from the size and internal connectivity of these bags of words, this does not shed light on their temporal distribution. In a recent study [SNM+15, SNM+14], we analyse whether the volume of collaborations of one author together with the relevance of his collaborators is somewhat related to his research performance over time. Although this study focuses on the dynamic patterns of network interactions amongst authors and their scholarly evolution, the techniques and the methodologies applied to handle data longitudinally might be useful to discern temporary interests from those that are persistent in users' online publications. Also in this scope, the possibility of predicting users' future interests from their current ones and any other convenient information (contacts' interests, current trends, etc.) might be explored in future research.

On the other hand, the high adoption rate of sensor-rich, Internet-enabled, mobile devices has allowed the emergence and massive adoption of Location-Based Social Networks, a new concept of social media that allows users to "check in" at physical places and share the location with their online friends, bridging the gap between the real and online worlds. Thus, an ongoing branch of enquiry might concern how to **include this location data in the intermediary model** presented in this dissertation. That is, ties and contexts vary with time, but they also can vary depending on users' location. Besides, the combination of social spheres with data about users' location would surely be useful to improve the creation of routes in opportunistic networks [PPC06]. That is, given the growing privacy concern in opportunistic networks [PH11, ZR12], not all nodes will be willing to transfer their resources to serve as intermediary nodes in the network. However, this would surely change if these nodes that wish to communicate were individuals with whom they have a connection in some social media – trusted individuals. This would involve not personalising online services, but taking the spheres out from the bubble which social media has become to benefit services in the *real* – as opposed to online – world.

The social contexts model explored in Chapter 3 is based on the idea that users talk about topics that are interesting to them but, with the analysis made so far, we cannot determine **users' opinions** about these interests. Therefore, it remains to investigate the inference of users' opinions towards these topics, and

specifically to determine whether the expressed opinion in content dealing with a given topic is positive, negative, or neutral [PL08]. It may be interesting to analyse users' personal feelings, views or beliefs expressed in this user generated content in order to provide a more accurate view of users' preferences and opinions in our social spheres model.

Finally, there are other technological applications of our model of social spheres besides those that we showed in Chapter 5 that could form the focus of further research. For example, in Section 5.3 we proposed a socially-enhanced application that helps users gain attention in social media by selecting the topics to discuss on social media sites. However, accurately selecting the piece of content to publish is not a task exclusive to social media users, but large providers of content have also to identify the most relevant and useful pieces to show to their users and win their attention. In our recent work [SHA15], we applied an attention economy solution to **generate the most informative content for the users** of a social media site with the aim of keeping and even increasing users' engagement with the site. Although this solution works independently of users' preferences, opinions, etc., i.e. takes into account only the content to determine its relevance for all the users in the system, it seems likely that social spheres do affect the relevance of the content. That is, that given the similarity between *contextualised social spheres* and the content, this content is relevant for only some subset of users in the system, with different degrees of relevance.

Bibliography

[adL] AdLemons. `http://adlemons.com`. [Online; accessed 06-March-2015].

[AGHT11] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on Twitter for personalized news recommendations. In *User Modeling, Adaption and Personalization*, pages 1–12. Springer, 2011.

[AHH+13] Fabian Abel, Eelco Herder, Geert-Jan Houben, Nicola Henze, and Daniel Krause. Cross-system user modeling and personalization on the Social Web. *User Modeling and User-Adapted Interaction (UMUAI), Special Issue on Personalization in Social Web Systems*, 23(2-3):169–209, 2013.

[AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[AZ12] Charu C. Aggarwal and Cheng-Xiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[BB07] David Beer and Roger Burrows. Sociology and, of and in Web 2.0: Some initial considerations. *Sociological Research Online*, 12(5):17, 2007.

[BBK+11] Lars Backstrom, Eytan Bakshy, Jon Kleinberg, Thomas M. Lento, and Itamar Rosenn. Center of Attention: How Facebook Users Al-

locate Attention Across Friends. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, ICWSM '11, 2011.

[BCD+09] Nilanjan Banerjee, Dipanjan Chakraborty, Koustuv Dasgupta, Sumit Mittal, Anupam Joshi, Seema Nagar, Angshu Rai, and Sameer Madan. User interests in social media sites: An exploration with micro-blogs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1823–1826, New York, NY, USA, 2009. ACM.

[BDS+14] Mohamed Ben-Khalifa, Rebeca P. Díaz-Redondo, Sandra Servia-Rodríguez, Ana Fernández-Vilas, and Rafael López-Serrano. Is There a Crowd? Experiences in using Density-Based Clustering and Outlier Detection. In *International Conference on Mining Intelligence and Knowledge Exploration (MIKE)*, Cork, Ireland, December 2014.

[BGL10] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *43rd Hawaii International Conference on System Sciences (HICSS)*, pages 1–10. IEEE, 2010.

[Bir07] William F. Birdsall. Web 2.0 as a social movement. *Webology*, 4(2):5–11, 2007.

[BK13] Moira Burke and Robert Kraut. Using facebook after losing a job: Differential benefits of strong and weak ties. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1419–1430, New York, NY, USA, 2013. ACM.

[BL11] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644. ACM, 2011.

[BLB+12] Jack F. Bravo-Torres, Martín López-Nores, Yolanda Blanco-Fernández, Sandra Servia-Rodríguez, and Jorge García-Duque. A virtualization layer for mobile consumer devices to support demanding

communication services in vehicular ad-hoc networks. In *IEEE International Conference on Consumer Electronics (ICCE)*, pages 225–226, Las Vegas, USA, January 2012.

[BLPM12] Yolanda Blanco-Fernández, Martín López-Nores, José J. Pazos-Arias, and Manuela I. Martín-Vicente. Spreading Influence Values over Weighted Relationships among Users of Several Social Networks. In *International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 149–154. IEEE, 2012.

[BMPZ11] John W Byers, Michael Mitzenmacher, Michalis Potamias, and Georgios Zervas. A Month in the Life of Groupon. *arXiv preprint arXiv:1105.0903*, 2011.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[Bon01] Monica Bonett. Personalization of web services: opportunities and challenges. *Ariadne*, 28, 2001.

[BRMA12] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 519–528, New York, NY, USA, 2012. ACM.

[BSH+10] Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, and Ed H. Chi. Eddi: Interactive Topic-based Browsing of Social Status Streams. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 303–312, New York, NY, USA, 2010. ACM.

[Bur02] Robin Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[CCPP04] Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of*

the International Conference on Language Resources and Evaluation, LREC '04, 2004.

[CDCS10] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[CF10] Wei Chen and S. Fong. Social network collaborative filtering framework and online trust factors: A case study on Facebook. In *Proceedings of the International Conference on Digital Information Management*, ICDIM, pages 266–273. IEEE, 2010.

[CNC11] Jilin Chen, Rowan Nairn, and Ed Chi. Speak little and well: Recommending conversations in online social streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 217–226, New York, NY, USA, 2011. ACM.

[CPV01] Corinna Cortes, Daryl Pregibon, and Chris Volinsky. Communities of interest. In Frank Hoffmann, David J. Hand, Niall Adams, Douglas Fisher, and Gabriela Guimaraes, editors, *Advances in Intelligent Data Analysis*, volume 2189 of *Lecture Notes in Computer Science*, pages 105–114. Springer Berlin Heidelberg, 2001.

[CV07] Rudi L. Cilibrasi and Paul M.B. Vitanyi. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[DDL+90] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by Latent Semantic Analysis. *JAsIs*, 41(6):391–407, 1990.

[DFPS12] Rebeca P. Díaz-Redondo, Ana Fernández-Vilas, José J. Pazos-Arias, and Sandra Servia-Rodríguez. A Social P2P Approach for Personal Knowledge Management in the Cloud. In *On the Move to Meaningful Internet Systems: OTM 2012 Workshops*, volume 7567 of *Lecture*

Notes in Computer Science, pages 585–594, Rome, Italy, September 2012. Springer Berlin Heidelberg.

[DOMA12] Anton Dimitrov, Alexandra Olteanu, Luke Mcdowell, and Karl Aberer. Topick: Accurate Topic Distillation for User Streams. In *Proceedings of the International Conference on Data Mining Workshops (ICDMW)*, pages 882–885, 2012.

[DPH12] Juang-Lin Duan, Shashi Prasad, and Jen-Wei Huang. Discovering Unknown But Interesting Items on Personal Social Network. 7302:145–156, 2012.

[dro] Dropbox. `https://www.dropbox.com`. [Online; accessed 06-March-2015].

[Dun98] Robin I.M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology*, 6:178–190, 1998.

[faca] Facebook. `https://www.facebook.com`. [Online; accessed 06-March-2015].

[Facb] Facebook Ads. `https://www.facebook.com/business/products/ads`. [Online; accessed 06-March-2015].

[Facc] Facebook Buttons. `https://developers.facebook.com/docs/plugins`. [Online; accessed 06-March-2015].

[Facd] Facebook Questions. `https://www.facebook.com/help/182071178590498`. [Online; accessed 06-March-2015].

[Face] Facebook statistics. `http://newsroom.fb.com/company-info/`. [Online; accessed 06-March-2015].

[FD07] Brendan J. Frey and Delbert Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, 2007.

[FDS14] Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and Sandra Servia-Rodríguez. IPTV parental control: A collaborative model for the Social Web. *Information Systems Frontiers*, pages 1–16, 2014.

[Fri80] Noah Friedkin. A test of structural features of Granovetter's strength of weak ties theory. *Social Networks*, 2(4):411–422, 1980.

[FS07] Ronen Feldman and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.

[GEVL13] Rebecca Gray, Nicole B. Ellison, Jessica Vitak, and Cliff Lampe. Who wants to know?: Question-asking and answering practices among facebook users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1213–1224, New York, NY, USA, 2013. ACM.

[GH06] Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[GK09] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220. ACM, 2009.

[GKBM11] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. Practical Recommendations on Crawling Online Social Networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1872–1892, 2011.

[GKF+06] Scott Garriss, Michael Kaminsky, Michael J. Freedman, Brad Karp, David Mazières, and Haifeng Yu. Re: reliable email. In *Proceedings of the conference on Networked Systems Design & Implementation*, NSDI '06, pages 22–22, 2006.

[GL12] Pritam Gundecha and Huan Liu. Mining social media: a brief introduction. *Tutorials in Operations Research*, 1(4), 2012.

[GM07] Evgeniy Gabrilovich and Shaul Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI '07, pages 1606–1611, 2007.

[Gof59] Erving Goffman. The presentation of self in everyday life. 1959.

[Gra73] Mark S. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

[Gra83] Mark S. Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, 1(1):201–233, 1983.

[Gra95] Mark Granovetter. *Getting a job: A study of contacts and careers*. University of Chicago Press, 1995.

[GRM+12] Przemyslaw A. Grabowicz, José J. Ramasco, Esteban Moro, Josep M. Pujol, and Victor M. Eguiluz. Social features of online networks: The strength of intermediary ties in online social media. *PloS ONE*, 7(1):e29358, 2012.

[gro] Groupon. `http://www.groupon.com`. [Online; accessed 06-March-2015].

[GXL+10] Mike Gartrell, Xinyu Xing, Qin Lv, Aaron Beach, Richard Han, Shivakant Mishra, and Karim Seada. Enhancing Group Recommendation by Incorporating Social Relationship Interactions. In *Proceedings of the ACM International Conference on Supporting Group Work*, GROUP '10, pages 97–106, New York, NY, USA, 2010. ACM.

[GZR+10] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. Social Media Recommendation Based on People and Tags. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 194–201, New York, NY, USA, 2010. ACM.

[Ham10] Eran Hammer-Lahav. RFC 5849: The OAuth 1.0 protocol, 2010.

[HD10] Liangjie Hong and Brian D. Davison. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[HH09] Courtenay Honey and Susan C. Herring. Beyond Microblogging: Conversation and Collaboration via Twitter. In *Proceedings of the Hawaii*

International Conference on System Sciences, HICSS '09, pages 1–10. IEEE, 2009.

[HMI03] Keiichiro Hoashi, Kazunori Matsumoto, and Naomi Inoue. Personalization of User Profiles for Content-based Music Retrieval Based on Relevance Feedback. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, pages 110–119, New York, NY, USA, 2003. ACM.

[Hof99] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57. ACM, 1999.

[Hoo] HootSuite. `http://hootsuite.com/`. [Online; accessed 06-March-2015].

[HPV06] Shawndra Hill, Foster Provost, and Chris Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276, 2006.

[HRW09] Bernardo Huberman, Daniel Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), 2009.

[Hug11] Ulrike Hugl. Reviewing person's value of privacy of online social networking. *Internet Research*, 21(4):384–407, 2011.

[JD88] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[JMF99] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999.

[JWL+11] Xin Jin, Chi Wang, Jiebo Luo, Xiao Yu, and Jiawei Han. LikeMiner: A System for Mining the Power of 'Like' in Social Media Networks. In *Proceedings of the ACM SIGKDD International Conference*

on Knowledge Discovery and Data Mining, KDD '11, pages 753–756, New York, NY, USA, 2011. ACM.

[KGA08] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A Few Chirps About Twitter. In *Proceedings of the First Workshop on Online Social Networks*, WOSN '08, pages 19–24, New York, NY, USA, 2008. ACM.

[KH10] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59 – 68, 2010.

[KLP10] Jeon Hyung Kang, Kristina Lerman, and Anon Plangprasopchok. Analyzing Microblogs with Affinity Propagation. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 67–70, New York, NY, USA, 2010. ACM.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web*, WWW '10, pages 591–600, 2010.

[KN09] Indika Kahanda and Jennifer Neville. Using Transactional Information to Predict Link Strength in Online Social Networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, ICWSM, 2009.

[KPV14] Márton Karsai, Nicola Perra, and Alessandro Vespignani. Time varying networks and the weakness of strong ties. *Scientific Reports*, 4, 2014.

[KR09] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[KS97] Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. Technical Report 1997-75, Stanford InfoLab, February 1997. Previous number = SIDL-WP-1997-0059.

[LD86] Nan Lin and Mary Dumin. Access to occupations through social ties. *Social networks*, 8(4):365–385, 1986.

[LWD+12] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *KDD*, pages 1023–1031, 2012.

[LZH08] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720, 2008.

[MB11] Alice E. Marwick and Danah Boyd. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133, 2011.

[MGR+12] Manuela I. Martín-Vicente, Alberto Gil-Solla, Manuel Ramos-Cabrer, Yolanda Blanco-Fernández, and Sandra Servia-Rodríguez. Semantics-driven recommendation of coupons through Digital TV: Exploiting synergies with social networks. In *IEEE International Conference on Consumer Electronics (ICCE)*, pages 564–565, Las Vegas, USA, January 2012.

[MLdlR] Miquel Montaner, Beatriz López, and Josep Lluís de la Rosa. A taxonomy of recommender agents on the internet. *Artificial Intelligence Review*, 19(4):285–330.

[MMS06] Bernd Marcus, Franz Machilek, and Astrid Schütz. Personality in cyberspace: personal web sites as media for personality expressions and impressions. *Journal of personality and social psychology*, 90(6):1014, 2006.

[MPC10] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.

[MSB+14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP

natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[MSLC01] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[Mut04] Paul Mutton. Inferring and visualizing social networks on internet relay chat. In *Eighth International Conference on Information Visualisation*, IV '04, pages 35–43. IEEE, 2004.

[Nad57] Siegfried F. Nadel. *The theory of social structure*. Cohen & West London, 1957.

[NSV11] Meena Nagarajan, Amit Sheth, and Selvam Velmurugan. Citizen sensor data mining, social media analytics and development centric web applications. In *Proceedings of the 20th international conference companion on World Wide Web*, pages 289–290. ACM, 2011.

[OKA10] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, ICWSM '10, 2010.

[OS05] John O'Donovan and Barry Smyth. Trust in Recommender Systems. In *Proceedings of the International Conference on Intelligent User Interfaces*, IUI '05, pages 167–174, New York, NY, USA, 2005. ACM.

[OSH+07] Jukka-Pekka Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and Albert-László Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[Pap03] Mike P Papazoglou. Service-oriented computing: Concepts, characteristics and directions. In *Proceedings of the Fourth International Conference on Web Information Systems Engineering (WISE)*, pages 3–12. IEEE, 2003.

[PG11] Marco Pennacchiotti and Siva Gurumurthy. Investigating topic models for social media user recommendation. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 101–102, New York, NY, USA, 2011. ACM.

[PH11] Iain Parris and Tristan Henderson. The impact of location privacy on opportunistic networks. In *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–6, June 2011.

[PL08] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, january 2008.

[PMK12] Katrina Panovich, Rob Miller, and David Karger. Tie strength in question & answer on social network sites. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1057–1066, New York, NY, USA, 2012. ACM.

[PPC06] Luciana Pelusi, Andrea Passarella, and Marco Conti. Opportunistic Networking: Data Forwarding in Disconnected Mobile Ad Hoc Networks. *Communications Magazine, IEEE*, 44(11):134–141, November 2006.

[PPM04] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet::Similarity: Measuring the Relatedness of Concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[PPZ+12] Diana Palsetia, Md M.A. Patwary, Kunpeng Zhang, Kathy Lee, Christopher Moran, Yves Xie, Daniel Honbo, Ankit Agrawal, Weikeng Liao, and Alok Choudhary. User-Interest based Community Extraction in Social Networks. In *Proceedings of the KDD workshop on Social Network Mining and Analysis (SNAKDD)*, 2012.

[PRS11] Srinivasan Parthasarathy, Yiye Ruan, and Venu Satuluri. Community Discovery in Social Networks: Applications, Methods and Emerging

Trends. In Charu C. Aggarwal, editor, *Social Network Data Analytics*, pages 79–113. Springer US, 2011.

[PTDL08] Michael P Papazoglou, Paolo Traverso, Schahram Dustdar, and Frank Leymann. Service-oriented computing: a research roadmap. *International Journal of Cooperative Information Systems*, 17(02):223–255, 2008.

[PVDH07] Mike P Papazoglou and Willem-Jan Van Den Heuvel. Service oriented architectures: approaches, technologies and research issues. *The VLDB journal*, 16(3):389–415, 2007.

[QAC12] Daniele Quercia, Harry Askham, and Jon Crowcroft. TweetLDA: supervised topic classification and link prediction in Twitter. In *Proceedings of the Annual ACM Web Science Conference*, WebSci '12, New York, NY, USA, 2012. ACM.

[Rai09] Geoffrey Raines. Cloud computing and soa. *MITRE technical papers, MITRE Corp., Massachusetts, USA*, 2009.

[RKT11] Aniket Rangrej, Sayali Kulkarni, and Ashish V. Tendulkar. Comparative study of clustering techniques for short text documents. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 111–112, New York, NY, USA, 2011. ACM.

[Rou87] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53–65, 1987.

[RSvZ10] Adam Rae, Börkur Sigurbjörnsson, and Roelof van Zwol. Improving tag recommendation using social networks. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 92–99, Paris, France, France, 2010. The Centre de Hautes Etudes Internationales d'Informatique Documentaire.

[SC11] Amit Sharma and Dan Cosley. Network-Centric Recommendation: Personalization with and in Social Networks. In *IEEE 3rd International Conference on Privacy, Security, Risk and Trust (PASSAT)*

and IEEE 3rd International Conference on Social Computing (SocialCom), pages 282–289, 2011.

[SDBA12] Alistair Sutcliffe, Robin I.M. Dunbar, Jens Binder, and Holly Arrow. Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology*, 103(2):149–168, 2012.

[SDF+14] Sandra Servia-Rodríguez, Rebeca P. Díaz-Redondo, Ana Fernández-Vilas, Yolanda Blanco-Fernández, and José J. Pazos-Arias. A tie strength based model to socially-enhance applications and its enabling implementation: mySocialSphere. *Expert Systems with Applications*, 41(5):2582 – 2594, 2014.

[SDF15] Sandra Servia-Rodríguez, Rebeca P. Díaz-Redondo, and Ana Fernández-Vilas. Are tweets biased by audience? an analysis from the view of topic diversity. In *International Social Computing, Behavioral-Cultural Modeling and Prediction Conference (SBP'15)*, Lecture Notes in Computer Science, Washington D.C., USA, April 2015.

[SDFP12] Sandra Servia-Rodríguez, Rebeca P. Díaz-Redondo, Ana Fernández-Vilas, and José J. Pazos-Arias. Using Facebook activity to infer social ties. In *International Conference on Cloud Computing and Services Science, CLOSER*, Porto, Portugal, April 2012.

[SDFP13] Sandra Servia-Rodríguez, Rebeca P. Díaz-Redondo, Ana Fernández-Vilas, and José J. Pazos-Arias. Mining Facebook Activity to Discover Social Ties: Towards a Social-Sensitive Ecosystem. In *Cloud Computing and Services Science*, volume 367 of *Communications in Computer and Information Science*, pages 71–85. Springer, 2013.

[SFDP12] Sandra Servia-Rodríguez, Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and José J. Pazos-Arias. Inferring Ties for Social-Aware Ambient Intelligence: The Facebook Case. In *International Symposium on Ambient Intelligence (ISAMI)*, volume 153 of *Advances in Intelligent and Soft Computing*, pages 75–83, Salamanca, Spain, March 2012. Springer Berlin Heidelberg.

[SFDP13a] Sandra Servia-Rodríguez, Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and José J. Pazos-Arias. Comparing Tag Clustering Algorithms for Mining Twitter Users' Interests. In *International Conference on Social Computing (SocialCom)*, pages 679–684, Washington D.C., USA, September 2013.

[SFDP13b] Sandra Servia-Rodríguez, Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and José J. Pazos-Arias. Inferring Contexts from Facebook Interactions: A Social Publicity Scenario. *IEEE Transactions on Multimedia*, 15(6):1296–1303, October 2013.

[SHA15] Sandra Servia-Rodríguez, Bernardo A. Huberman, and Sitaram Asur. Deciding what to display: maximizing the information value of social media. In *Workshop on Modeling and Mining Temporal Interactions (M2TI) at ICWSM' 15*, Oxford, UK, May 2015.

[SKKR01] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[SKQ09] Ross Shannon, Eugene Kenny, and Aaron Quigley. Using Ambient Social Reminders to Stay in Touch with Friends. *International Journal of Ambient Computing and Intelligence (IJACI)*, 1(2):70–78, 2009.

[SNM+14] Sandra Servia-Rodríguez, Anastasios Noulas, Cecilia Mascolo, Ana Fernández-Vilas, and Rebeca P. Díaz-Redondo. The evolution of your success lies in the centre of your co-authorship network. In *Quantifying Success (2.0) –co-located with ECCS 2014*, Lucca, Italy, September 2014.

[SNM+15] Sandra Servia-Rodríguez, Anastasios Noulas, Cecilia Mascolo, Ana Fernández-Vilas, and Rebeca P. Díaz-Redondo. The evolution of your success lies at the centre of your co-authorship network. *PLoS ONE*, 10:e0114302, 03 2015.

[SP06] Michael Strube and Simone P. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the National*

Conference on Artificial Intelligence, volume 21 of *AAAI '06*, pages 1419–1424, 2006.

[SPUP02] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 253–260, New York, NY, USA, 2002. ACM.

[SQV+14] Quan Z Sheng, Xiaoqiang Qiao, Athanasios V Vasilakos, Claudia Szabo, Scott Bourne, and Xiaofei Xu. Web services composition: A decade's overview. *Information Sciences*, 280:218–238, 2014.

[SRF13] Bracha Shapira, Lior Rokach, and Shirley Freilikhman. Facebook single and cross domain data for recommendation systems. *User Modeling and User-Adapted Interaction*, 23(23):211–247, 2013.

[TDH05] Jaime Teevan, Susan T Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456. ACM, 2005.

[TQKH12] Xin Tan, Li Qin, Yongbeom Kim, and Jeffrey Hsu. Impact of privacy concern in social networking web sites. *Internet Research*, 22(2):211–233, 2012.

[Twe] TweetDeck. `https://about.twitter.com/products/tweetdeck`. [Online; accessed 06-March-2015].

[TWH05] J.R. Tyler, D.M. Wilkinson, and Bernardo A. Huberman. E-mail as Spectroscopy: Automated Discovery of Dommunity Structure within Organizations. *The Information Society*, 21(2):143–153, 2005.

[twia] Twitter. `https://twitter.com`. [Online; accessed 06-March-2015].

[Twib] Twitter Buttons. `https://about.twitter.com/resources/buttons`. [Online; accessed 06-March-2015].

[VMCG09] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the Evolution of User Interaction in Facebook. In *Proceedings of the ACM Workshop on Online Social Networks*, WOSN '09, pages 37–42, New York, NY, USA, 2009. ACM.

[vR79] Cornelis Joost van Rijsbergen. Information retrieval. 1979.

[WB10] Yi Wei and M. Brian Blake. Service-Oriented Computing and Cloud Computing: Challenges and Opportunities. *Internet Computing*, 14(6):72–75, 2010.

[WBS+09] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. User Interactions in Social Networks and Their Implications. In *Proceedings of the ACM European Conference on Computer Systems*, EuroSys '09, pages 205–218, New York, NY, USA, 2009. ACM.

[Whi08] Harrison C. White. *Identity and Control: How Social Formation Emerge*. Princeton University Press, 2008.

[WLJH10] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[WM08] Ian H. Witten and David Milne. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30, 2008.

[WTC09] Chen Wen, Bernard CY Tan, and Klarissa Ting-Ting Chang. Advertising effectiveness on social network sites: an investigation of tie strength, endorser expertise and product type on consumer purchase intention. In *Proceedings of the International Conference on Information Systems*, ICIS '09, page 151, 2009.

[XNR10] Rongjing Xiang, Jennifer Neville, and Monica Rogati. Modeling Relationship Strength in Online Social Networks. In *Proceedings of the*

19th International Conference on World Wide Web, WWW '10, pages 981–990, New York, NY, USA, 2010. ACM.

[YKGF06] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. Sybilguard: Defending against sybil attacks via social networks. *SIGCOMM Comput. Commun. Rev.*, 36(4):267–278, 2006.

[You] Youtube Ads. `http://www.youtube.com/yt/advertise/`. [Online; accessed 06-March-2015].

[ZCB10] Qi Zhang, Lu Cheng, and Raouf Boutaba. Cloud computing: state-of-the-art and research challenges. *Journal of Internet Services and Applications*, 1(1):7–18, 2010.

[ZJC11] Mimi Zhang, BernardJ. Jansen, and Abdur Chowdhury. Business engagement on twitter: a path analysis. *Electronic Markets*, 21(3):161–175, 2011.

[ZR12] Sameh Zakhary and Milena Radenkovic. Utilizing social links for location privacy in opportunistic delay-tolerant networks. In *2012 IEEE International Conference on Communications (ICC)*, pages 1059–1063, June 2012.

[ZWF+12] Jichang Zhao, Junjie Wu, Xu Feng, Hui Xiong, and Ke Xu. Information propagation in online social networks: a tie-strength perspective. *Knowledge and Information Systems*, 32(3):589–608, 2012.

[ZYL+12] Xiaojian Zhao, Jin Yuan, Guangda Li, Xiaoming Chen, and Zhoujun Li. Relationship strength estimation for online social networks with the study on Facebook. *Neurocomputing*, 95(0):89 – 97, 2012.

Resumen

A. Motivación

La aparición de la Web a principios de los años 90 ha cambiado forma en la que las personas se relacionan, superándose las limitaciones impuestas por el mundo físico y permitiendo comunicarse a miles de kilómetros de distancia. Inicialmente pensada para facilitar el intercambio de información entre físicos nucleares, la Web ha evolucionado hasta convertirse en un medio para comunicarse, encontrar información e incluso para entretenerse. Aunque ya desde sus inicios han existido iniciativas para convertir la Web en un entorno más social, no fue hasta el estallido de la Web Social a principios de los años 2000 cuando se produjo la socialización de la Web a través de la involucración y participación activa de los usuarios, ya que estos empezaron a actuar no sólo como típicos consumidores, si no también como creadores de contenidos. Estas tecnologías han proporcionado poderosas herramientas para difundir información objetiva, opiniones y en definitiva cualquier tipo de contenido que se desee compartir con los círculos sociales, desarrollándose nuevas formas de comunicación que van más allá de las típicas interacciones en persona. Dentro de esta definición general, existen distintos tipos de medios sociales que han nacido al abrigo de la Web Social: blogs (Blogger, WordPress, ...), redes sociales (Facebook, Foursquare, ...), proyectos colaborativos (Wikipedia, OpenStreetMap, ...), comunidades de contenido (YouTube, LastFm, ...), etc. Aunque no existe un procedimiento sistemático para categorizar dichos sitios, algunos investigadores como Kaplan y Haenlein [KH10] han propuesto una categorización basada en distintas teorías basadas en medios de comunicación y procesos sociales. Además, incluso dentro

de cada tipo de medio social existen diferencias en cuanto a los mecanismos que los usuarios de unos sitios u otros disponen para compartir contenido: tipo de contenido a compartir (texto, fotos, vídeos, ...), visibilidad del contenido (visible por todos, sólo por los amigos, ...), propósito y ámbito del sitio (personal, profesional, ...), etc. Con independencia de la tecnología social considerada, la información de este contenido compartido, junto con otros metadatos como el tiempo o la localización, los convierten en valiosas fuentes de datos que reflejan los intereses de los usuarios, sus opiniones, etc [GL12]. El uso generalizado de estas tecnologías (en septiembre de 2014, Facebook, la red social *online* más popular, contaba con más de un billón de usuarios activos en el último mes [Face]) ha supuesto la disponibilidad de grandes cantidades de este contenido dinámico y continuamente actualizado generado por los usuarios, cuyo análisis beneficia a distintas aplicaciones en diversos ámbitos, desde el de los negocios al de las ciencias sociales [BL11, KH10, GK09, BGL10].

La personalización, es decir, el uso de tecnología para adaptarse a las diferencias entre distintos individuos, ha jugado un papel importante en el éxito de los servicios *online*. Durante años, la personalización ha supuesto que las aplicaciones se encargasen de obtener información de sus usuarios, la cual, después de ser cuidadosamente analizada, se utilizaba para mostrarles contenido adecuado a sus preferencias. Tradicionalmente, dicha información se obtenía a partir de un histórico de sesiones previas, o a través de interacciones en tiempo real [Bon01]. Pero esta forma de proceder presenta diversos inconvenientes y limitaciones principalmente relacionados con (i) la ausencia de información cuando un nuevo servicio se lanza por primera vez – *cold-start problem* [SPUP02] – y (ii) que incluso los usuarios más activos sólo han puntuado un pequeño subconjunto de todos los servicios disponibles, lo que hace que los datos sean dispersos e insuficientes para identificar similitudes entre los intereses de los usuarios – *sparsity problem* [SKKR01]. Recientemente, la aparición de la Web Social y su gran popularidad han transformado la Web en un entorno inundado de contenido creado por los usuarios. La gran aceptación de estas tecnologías, su penetración en todos los sectores sociales y la libertad con la que los usuarios participan en ella sugieren que utilizar dicho contenido con propósitos de personalización beneficiaría enormemente a los servicios [BB07, Bir07], permitiéndoles incluso evitar problemas como el *cold-start* o la *sparsity*. En su objetivo de satisfacer a sus

usuarios, los servicios necesitan conocer sus intereses, qué les gusta. La mayoría de los sitios sociales permiten a sus usuarios crear perfiles, los cuales normalmente incluyen datos demográficos y geográficos e incluso intereses, y el simple análisis de esta información se ha demostrado suficiente para desarrollar potentes servicios personalizados. Ejemplo de ello son los servicios de publicidad de Facebook [Facb] y Youtube [You]. Los investigadores también han demostrado la utilidad de analizar el contenido espontáneo que los usuarios publican inconscientemente en sus cuentas para crear experiencias emocionantes que animen al usuario a continuar utilizando los servicios [TDH05, JWL+11, AGHT11].

La Web Social, y especialmente las redes sociales *online*, también han provisto a los usuarios de un medio de comunicación *online*, y por tanto de una oportunidad para la aparición y fortalecimiento de relaciones *online*. Esto, junto con la tendencia de los individuos a asociarse y establecer vínculos con individuos similares (*homofilia*) [MSLC01], sugiere que considerar los intereses de los amigos del usuario sería beneficioso para la personalización de los servicios. Aunque muchas tecnologías de la Web Social, y especialmente las redes sociales, permiten a los individuos *conectarse* con otros, no todas estas conexiones suponen la existencia de una relación social. Al contrario, una persona sólo puede mantener un número limitado de relaciones e incluso un número limitado de ellas a distintos niveles de cercanía [Dun98]. Por tanto, se necesita distinguir las relaciones reales de simples comunicaciones/interacciones esporádicas si se desean desarrollar *servicios socialmente mejorados* de forma eficiente.

Casi a la vez que lo anterior, otro fenómeno revolucionario ha aparecido con fuerza en el panorama tecnológico: el llamado paradigma de Computación Orientado a Servicios (SOC) [Pap03]. Este paradigma de computación ha cambiado la forma en la que las aplicaciones software son diseñadas, distribuidas y consumidas. En SOC, los servicios se utilizan como bloques básicos para construir aplicaciones rápidas, de bajo coste, seguras y fiables, reduciendo la necesidad de desarrollar nuevos componentes software cada vez que aparece un nuevo proceso de negocio [PVDH07, PTDL08]. Mediante la utilización de lenguajes de descripción estándar, un servicio puede exponer su interfaz al mundo exterior para ser descubierto e invocado de forma aislada o a través de la composición de múltiples servicios. Como ejemplos destacados de esta nueva tendencia,

compañías como Google, Amazon, Twitter y Facebook ofrecen servicios Web para acceder a algunos de sus recursos, permitiendo a aplicaciones de terceros combinar y reutilizar sus servicios [SQV+14]. La inevitable penetración de SOC y el paradigma de Computación en la Nube [WB10, Rai09] con el que está estrechamente relacionado – donde el llamado *Everything as a Service (EaaS o XaaS)* permite una Nube (metáfora de Internet) que almacena recursos que serán distribuidos como servicios de una elevada granularidad y que pueden ser compuestos de una forma flexible en respuesta a necesidades complejas [ZCB10] – sugiere la necesidad de aplicaciones que sean desarrolladas de tal forma que puedan ser integradas en servicios existentes y/o construidas sobre ellos.

Lo indicado anteriormente nos lleva a que la personalización de los servicios *online* no debería mantenerse al margen de esta tendencia. Es decir, la provisión de personalización debería ser proporcionada mediante servicios intermediarios para ser descubiertos e invocados por otros servicios que deseen ofrecer experiencias personalizadas a sus clientes sin que estos tengan que descubrir las preferencias de dichos usuarios. Además, los servicios encargados de ofrecer estas preferencias – perfiles – no deberían obviar el potencial del contenido que los usuarios comparten espontáneamente en la Web Social para obtener las preferencias de los usuarios y cualquier otra información que pueda ser de utilidad. Además, los individuos utilizan varios sitios sociales y la integración del contenido que generan en todos ellos es lo que caracteriza completamente su vida *online*. Para desarrollar tales servicios se necesitan resolver distintas cuestiones: cómo representar/modelar adecuadamente a los usuarios, cómo analizar adecuadamente el contenido generado por los usuarios para extraer información útil, cómo distinguir a los verdaderos amigos de simples conocidos, etc. En esta tesis nos centramos en estas cuestiones, y específicamente, en proponer y analizar distintas técnicas de análisis de datos para extraer información útil del contenido generado por los usuarios en distintos sitios sociales y representarlo de tal forma que pueda ser proporcionado a otros servicios para ser personalizados / mejorados socialmente de forma efectiva.

B. Tesis y contribuciones

La hipótesis explorada en esta tesis es la de que *un modelo intermediario del usuario construido mediante el análisis del contenido que éste ha generado en distintas plataformas de la Web Social puede ser explotado para crear o perfeccionar aplicaciones socialmente mejoradas.* Para examinar esta hipótesis, es necesario diseñar el modelo de tal forma que abarque los intereses del usuario y sus contactos, sin perder de vista el cómo este modelo podría ser utilizado por dichas aplicaciones tecnológicas.

Específicamente, en esta tesis se propone un modelo centrado en el usuario para personalizar servicios. Dicho modelo surge de aplicar distintas técnicas de minería de datos sobre el contenido que los usuarios comparten en distintas plataformas sociales con el objetivo de identificar sus intereses y la fortaleza de las relaciones que estos mantienen con otros usuarios. Dicho modelo proporciona distintas salidas o perfiles del usuario: *contextos sociales*, *esferas sociales* y *esferas sociales contextualizadas*. Con *contextos sociales* nos referimos a los temas en los que está interesado el usuario y que se obtienen analizando el contenido textual que éste ha publicado (y normalmente compartido con sus contactos), representados en forma de nubes de etiquetas. *Esferas sociales* es el término que acuñamos para referirnos al conjunto de usuarios con los que nuestro usuario interacciona junto con la fortaleza de su vínculo calculada a partir de sus interacciones. Finalmente, las *esferas sociales contextualizadas* surgen de la combinación de contextos y contactos. Cada una de estas esferas estaría formada por aquellos usuarios con los que nuestro usuario habla frecuentemente del tema del contexto de la esfera, junto con la fortaleza de su vínculo calculado teniendo en cuenta sólo aquellas interacciones en el ámbito del tema. En esta tesis nos centramos en detallar las técnicas que proponemos para obtener los contextos y esferas. También describimos cómo implementar este modelo de usuario en un prototipo de servicio encargado de monitorizar la actividad de los usuarios en los sitios sociales y de proporcionar los contextos y esferas para personalizar o mejorar socialmente un amplio número de aplicaciones y servicios, también descritos. Finalmente, y dada la sensibilidad de la información manejada y la creciente preocupación por la privacidad de los datos, tal servicio debería requerir

permiso a los usuarios tanto para acceder a sus datos en su nombre como para proporcionar sus contextos y/o esferas a otros servicios o aplicaciones.

Con todo esto, la contribución principal de esta investigación es la externalización de la provisión de personalización a los servicios *online* a través de la definición de un modelo de esferas sociales obtenido a partir de datos de interacción presentes en distintos sitios sociales. Para crear tales esferas, se proponen distintas estrategias:

1. Para descubrir y modelar los intereses de los individuos manifestados en la Web Social y aquellos contactos que comparen estos intereses con ellos a través del análisis del contenido que estos generan en los sitios sociales; y representar dichos *contextos sociales* mediante etiquetas de palabras representativas, lo que hace que puedan ser utilizados casi por cualquier aplicación,

2. Para medir y representar la fortaleza del vínculo entre dos individuos sociales desde la perspectiva de uno de ellos a través del análisis de evidencias de sus interacciones disponibles en los sitios sociales y obtenidas a través de *Interfaces de Programación de Aplicaciones – APIs –* (mensajes privados, retweets, menciones, etc.); y, al contrario que en otras propuestas previas, teniendo en cuenta distintos tipos de interacciones y contextos, el instante en el que las interacciones suceden, la gente involucrada en ellas y la frecuencia de las interacciones entre el usuario y el resto de sus contactos, y

3. Para mostrar cómo este modelo de esferas sociales puede ser utilizado para crear y mejorar servicios sociales tecnológicos tales como aplicaciones para obtener atención o encontrar expertos de confianza en la Web Social; y cómo dicho modelo puede ser fácilmente integrado en un servicio SOC que proporciona esferas y contextos a otros servicios bajo petición y siempre con el permiso de los usuarios.

C. Análisis y resultados

A lo largo de la tesis se explora la hipótesis de partida para lo cual se define y evalúa el modelo de esferas sociales. Se comienza proponiendo una metodología para obtener los intereses de los usuarios a partir de contenido textual que estos publican libremente en los sitios sociales (descripciones en las fotos, mensajes privados, ...). Dicha metodología tiene el potencial de ser desarrollada sin conocer, a priori, el número y la categoría de los intereses, ni ninguna otra información sobre los usuarios sobre los que estamos realizando la obtención. La metodología se basa en aplicar distintas técnicas de Procesado del Lenguaje Natural para obtener las palabras más representativas de este contenido que, junto con la relación semántica entre ellas, constituyen la personomía del usuario. Sobre esta personomía se aplica *clustering* para obtener clusters o grupos de palabras fuertemente relacionadas y finalmente, atendiendo a sus conversaciones con el usuario, los contactos se clasifican en los clusters o contextos. Para evaluar la efectividad de esta metodología se llevaron a cabo dos estudios distintos: (i) un estudio con usuarios en Facebook y (ii) otro estudio utilizando los hashtags de Twitter. Los resultados del primer estudio muestran que la consideración del contenido generado espontáneamente por los usuario permite predecir con gran precisión los ítems preferidos por los usuarios, demostrando así que esta metodología de obtención de intereses podría ser aplicada para mejorar las estrategias de publicidad social. Pero este estudio también ha evidenciado que los usuarios no son fiables cuando se les pregunta por sus intereses. El segundo estudio, el cual surge como un intento de superar las limitaciones ocasionadas por la poca fiabilidad de los participantes, así como de la escasez de datos del primero, reveló que (i) un algoritmo de clustering jerárquico (UPGMA) es el algoritmo que mejor se comporta en el ámbito de nuestra metodología y (ii) la idoneidad de utilizar la distancia Silhouette para seleccionar los parámetros de entrada a los algoritmos de clustering.

Esta investigación sobre la obtención y aplicación de los intereses de los usuarios a partir del contenido que estos generan en la Web Social con propósitos de personalización se basa en trabajos previos que tienen en cuenta la información de los usuarios, principalmente de sus perfiles, para personalizar aplicaciones

creadas en la misma plataforma en el que se obtiene el perfil. Pero nuestra contribución va un paso más adelante y, en vez de centrarse en el contenido estructurado de los perfiles, se basa en el análisis del contenido desestructurado que los usuarios publican libremente en los sitios web sociales. Además, en vez de imponer las limitaciones de una clasificación de intereses fijada a priori, nuestra solución se basa en una representación mediante nubes de etiquetas. Esto, además de suponer una elevada granularidad en la definición de intereses, hace que los contextos sociales puedan ser fácilmente integrados en casi cualquier aplicación o servicio que desee ser personalizado o mejorado socialmente.

Como segunda contribución al modelo de esferas sociales, se propone una medida del vínculo entre dos individuos centrada en el usuario y obtenida a partir de las evidencias de sus interacciones en la Web Social. Es decir, se describe la metodología para tasar la cercanía que un usuario percibe de su relación con otro individuo utilizando para ello evidencias de su actividad en diferentes plataformas de la Web Social (la fortaleza del vínculo entre dos usuarios, u y v, desde la perspectiva de u). La evaluación de esta medida o metodología se realizó mediante un estudio con usuarios reales en Facebook en el que (i) los participantes clasificaron a sus contactos en grupos de cercanía y (ii) dicha medida se utilizó para predecir el grupo en el que cada contacto había sido clasificado. Los resultados demostraron que esta medida clasifica aceptablemente a los contactos de los usuarios en círculos de cercanía y que los usuarios no son tan fiables como cabría esperar cuando se les pregunta acerca de las relaciones que mantienen con otros usuarios en Facebook.

Este estudio para medir la fortaleza de la relación entre dos individuos desde la perspectiva de uno de ellos se basa en trabajos previos centrados en tasar lo que Granovetter [Gra73] llamó fortaleza del vínculo entre dos individuos o *tie strength* en inglés. Los estudios iniciales en medir la fortaleza del vínculo se realizaron principalmente a través de encuestas, caracterizadas por proporcionar una visión estática y muy limitada. Pero ahora, gracias a la aparición de la Web Social y a su uso generalizado hay muchos más datos disponibles sobre interacciones entre usuarios lo que ha hecho que se factible la utilización de dichos datos para obtener la fortaleza del vínculo entre usuarios. Al contrario que trabajos previos centrados en analizar datos de interacción producidos en un

único sitio social [WBS+09, BRMA12, ZWF+12], nosotros tenemos en cuenta una visión más completa de la vida *online* de los usuarios, considerando evidencias de interacciones en todos las plataformas en los que dichos usuarios participan. Además, para tasar con precisión la cercanía entre dos individuos, tenemos en cuenta un amplio abanico de factores tales como distinguir entre diferentes tipos de interacciones y contextos, el tiempo en el que suceden las interacciones, la gente involucrada en ellas y la frecuencia de las interacciones entre los usuarios y el resto de sus contactos.

Finalmente en esta tesis también se describe cómo nuestro modelo de esferas sociales podría ser fácilmente integrado en un servicio *online* que siguiese el paradigma *software-as-a-service*, servicio que estaría encargado de construir, gestionar y entregar las esferas bajo petición y siempre con el permiso de los usuarios. El hecho de implementar el modelo de esferas sociales en un servicio que sigue el paradigma *software-as-a-service* y que proporciona soporte para otros servicios a través de una API REST incrementaría la disposición de las aplicaciones a considerar dicho servicio como una herramienta externa con la que proporcionar a sus usuarios experiencias personalizadas sin detrimento de su privacidad. También se analiza la posibilidad de usar las esferas sociales en dos aplicaciones socialmente mejoradas encargadas de ayudar a los usuarios a (i) obtener atención y (ii) encontrar expertos de confianza en la Web Social. Para el primero, se detallan los experimentos realizados para estudiar la relación entre la diversidad de temas hablados por los usuarios en sus publicaciones y el tamaño de su audiencia. El hallazgo de dicha relación demuestra la pertinencia de utilizar esferas sociales – contextos sociales – en una aplicación que alerte a los usuarios cuando estos deban diversificar o especializar los temas de sus publicaciones con el objetivo de incrementar su audiencia. Para el segundo, se detalla nuestra propuesta P2P social para ayudar a los individuos a encontrar información o servicios en la Web de forma efectiva. Dicha propuesta utiliza las esferas sociales – tanto la fortaleza de los vínculos como los contextos – para superar el aislamiento de los usuarios causado por su visión limitada de la Web Social – estos únicamente son conscientes de lo que ocurre en sus esferas sociales. Dado que estos no son los únicos servicios a mejorar socialmente, también se explican otras aplicaciones que podrían ser personalizadas utilizando esferas sociales.

La descripción en detalle del servicio basado en esferas sociales así como la relación de aplicaciones que podrían ser socialmente mejoradas con él constituyen la prueba definitiva de la viabilidad de usar el contenido generado por los usuarios en distintos medios sociales con propósitos de personalización. Además, la externalización de la provisión de personalización teniendo en cuenta el paradigma de computación orientado a objetos supone una reducción considerable de la carga de trabajo de los servicios, permitiendo a dichos servicios o aplicaciones centrarse únicamente en sus tareas principales. Finalmente, los usuarios son dueños de sus datos y son ellos quienes deciden compartirlos con aquellos servicios en los que confían, lo cual viene al auxilio de la creciente preocupación por la privacidad de los datos en los medios sociales.

Como conclusión, esta investigación proporciona evidencias de que es factible construir un modelo intermediario de usuario, tanto desde el punto de vista de sus intereses como de sus relaciones sociales, mediante el análisis del contenido que éste genera en los sitios sociales; y que dicho modelo puede ser utilizado para personalizar o mejorar socialmente un extenso número de aplicaciones sociales tecnológicas, tanto existentes o que puedan ser creadas, desde sistemas de recomendación a gestores de correo electrónico, por ejemplo.