Identification of User Interests in Social Media

by

Giray Havur

Submitted to the Graduate School of Sabancı University
in partial fulfillment of the requirements for the degree of
Master of Science

Sabanci University

August, 2014

APPROVED BY:

Assoc. Prof. Dr. Yücel Saygın .....
(Thesis Advisor)

Assoc. Prof. Dr. Berrin Yanıkoğlu .....

Assoc. Prof. Dr. Ali Koşar .....

DATE OF APPROVAL: .....

Identification of User Interests in Social Media

Giray Havur
Computer Science and Engineering, MS Thesis, 2014
Thesis Advisor: Yücel Saygın

Keywords: Online communities, information retrieval, semantic analysis, sentiment analysis, social networks, user ranking.

Abstract

Social media has taken an important part in our lives in a short amount of time. People share parts of their experiences, opinions, and interests with others in a timely-fashion on these platforms. In recent years, fast growth of user population in social media is not only driving the research towards analyzing its inhabitants for fulfilling their expectations but also making it a very crucial information source for decision making processes in societies and in businesses. In this work, we propose methods for identifying users and their interests by using the multimedia data shared in social media. We show effectiveness of these methods in three applications. Our first application considers extracting political interests of Turkish Twitter users. We collect tweets that include a set of predefined words representing two different political stances in Turkey. We extract profile images of the users who wrote those tweets and apply a computer vision technique called image context extraction on this set of images to obtain some textual explanations for each picture. The main goal of this work is inferring proportions of two different political stances to forecast results of March 2014 local elections. Our results show that the proportions obtained from our method are almost the same as the vote percentages of two parties. In our second application, we find Facebook profiles of people whose identification information (Name, surname and location) is given by querying Facebook Graph API. Each query result returns a number of profiles due to people having same name. We refine these results by checking location in profile pages online. Our method achieves a successful match rate of 88% (1332/1500 people). The third application deals with building a community about a given topic of interest by condensing existing communities in a social media platform. We collect members of the communities about the given topic in a set and apply our relevance scoring method on these members. Those who receive a score below a threshold value are assumed to be irrelevant to given topic and they are eliminated so that remaining users in the set are the ones relevant to given topic. We validated the results of our framework by a user-study. There is a 76% of match between user labelled and automated results.

Sosyal Medyada Kullanıcı İlgi Saptanımı

Giray Havur
Bilgisayar Bilimi ve Mühendisliği, Yüksek Lisans Tezi, 2014
Tez Danışmanı: Yücel Saygın

Anahtar Kelimeler: Çevrimiçi topluluklar, bilgi çekimi, semantik analiz, duygu analizi, sosyal ağlar, kullanıcı tasnifi.

Abstract

Sosyal medya kısa zamanda hayatlarımızda önemli bir yer edindi. İnsanlar deneyimlerini, fikirlerini ve ilgi alanlarını bu platform üzerinde gerçek zamanlı olarak paylaşmaya devam ediyorlar. Son yıllarda sosyal medyada süratle artan kullanıcı sayısı, bu alanda yapılan araştırmaların kullanıcı beklentilerini karşılamak üzere yoğunlaşmasına sebep olduğu gibi sosyal medyanın toplumlar ve iş dünyası içerisinde yer alan karar mekanizmalarında kritik bir önem taşımasına yol açtı. Bu çalışmamızda sosyal medya kullanıcılarını ve onların ilgi alanlarını saptamaya yönelik yöntemler geliştirdik. İlk uygulamamız, Türkiye'deki Twitter kullanıcılarının politik ilgi odağını belirlemeye yöneliktir. Ülkedeki iki farklı politik duruşu tarif eden iki anahtar kelime kümesi belirleyerek, bu kümelerdeki kelimeleri içeren tweetleri ve bu tweetlerin kullanıcılarını topluyoruz. Bu kullanıcıların profil fotoğrafları üzerinde bir bilgisayarlı görüntü işleme tekniği olan görüntü özüt çıkarımı yöntemini uyguluyor ve fotoğrafların içeriği hakkında metinsel bilgi elde ediyoruz. Bu bilgileri ve başta tanımladığımız anahtar kelime kümelerini kullanarak Mart 2014 yerel seçimlerinden önce iki farklı politik duruşu destekleyen grupların rakamsal oranlarını tahmin etmeyi hedefliyoruz. Uygulamamızda elde ettiğimiz sonuçlarla seçim sonuçlarının birebir örtüştüğünü gözlemledik. İkinci çalışmamızda insanların tanımlama bilgilerini (Örn. İsim, yaşadığı yer) kullanarak bu insanların Facebook profil sayfalarına ulaşmaya çalışıyoruz. 1500 tanımlama bilgisini kullanarak 1332 Facebook profiline ulaştık. Son uygulamamız ise bir ilgi alanı etrafında, bu konuyla ilgili insanları yakalamayı hedefliyor. Sosyal medya üzerinde halihazırda bu konuyla ilgili olan toplulukların üyelerini bir kümede toplayarak bu üyeleri önerdiğimiz ilgi değerlendirme yöntemiyle notlandırıyoruz. Eşik değerin altında notlandırılan üyeler, belirlenen ilgi alanıyla alakasız oldukları öngörülerek kümeden eleniyorlar ve böylece geride kalan üyeler belirlenen konuyla ilgili üyeler oluyor. Sonuçlarımızı doğrulamak için yaptığımız çalışmada, insanlar tarafından ilgili olduğu tespit edilen sosyal medya kullanıcılarıyla uygulamamızın tespit ettiği kullanıcılar arasında %76'lık bir örtüşme olduğunu gözlemledik.

"*Dancing all paths*

And swimming a song

around [t]his royal homeland

In the poorest full"

Acknowledgement

Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible.

My deepest gratitude is to my advisor, Dr. Yücel Saygın. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. His patience and support helped me overcome difficult times and finish this dissertation. I hope that one day I would become as good an advisor to my students as he has been to me.

Dr. Berrin Yanıkoğlu has been always there to listen and give advice which has been invaluable on both an academic and a personal level, for which I am extremely grateful. I am also thankful to her for supporting me and believing in me.

I am grateful to Dr. Dilek Tapucu for numerous discussions; for reading my reports, commenting on my views and helping me understand and enrich my ideas.

I would like to express my very sincere gratitude to Dr. Ali Koşar. I am indebted to him for his continuous encouragement and guidance.

Many friends have helped me stay sane through these difficult years. Peter Schüller, Damien Jade Duff, Ezgi Karakaş, Zeynep Sarıbatur, Zeynep Doğmuş, Mine Saraç, Soner Ulun, Gökay Çoruhlu, Ozan Tokatlı and Mustafa Yalçın. Your support and care helped me overcome setbacks and stay focused on my graduate study. I greatly value your friendship.

Most importantly, none of this would have been possible without the love and patience of my mother. I also would like to express my heart-felt gratitude to my grandparents Paulette and Yüksel and to my cousins Nebahet, Arif and Aslı. I always feel blessed to have you in my life. I have to give a special mention for my soul sisters and brothers: Ceren, Saniye, Doğa, Sinan, Barış, Ümit, Orkun and Can. My extended family has aided and encouraged me throughout this endeavor.

You are whom this thesis is dedicated to, has been a constant source of love, concern, support and strength all these years.

Contents

List of Figures

Chapter 1

Introduction

In recent years, social media is playing an important role in everyone's life. For instance, 82% of the world's online population has profiles in different social media and nearly 1 in every 5 minutes spent online is spent on social media sites nowadays [1]. Social media has become a part of our social lives as more and more people joined. These platforms provide us new ways for self expression and communication with other people.

There are many different types of social media in which users participate. Social networks, business networks, microblogging services, collaborative encyclopedias, virtual worlds, online multimedia sharing platforms are considered as social media services. Most of these services encourages their users for personalization, in other words, for elaborating their profiles. Users disclose their thoughts, emotions and interests by sharing multimedia content with other users.

As there is a great amount of public personal data in social media services, new research opportunities arose on analysis, modeling, and exploitation of user interests. In this thesis, we propose novel approaches towards user interest identification in different applications. We primarily work on the following research issues:

1. How to infer interests of social media users using multimedia data they publicly share

2. How to reach such users sharing a common interest

To solve the former problem, we integrate various methods from information retrieval, natural language processing and computer vision for estimating a quantitative score for user that indicate his/her interest to a topic. For example, we can identify user interests by analyzing his/her pictures such as profile image for extracting interests. An example to this method is our political party supporter analysis on Twitter detailed in Chapter 4.

To solve the latter problem, we take advantage of existing online communities in social media. We collected the users of online communities formed around an interest topic and refined this set of users by giving them a relevance score so that we select users with high scores. We propese a novel multi-stage computational framework using this approach in Chapter 6.

1.1 Contributions

Most work in this thesis are closely connected to real applications in social media. They are developed to addressing real world needs, therefore some of them can help improve user experiences in large scale social networking platforms such as Facebook, and bring light on user interests and preferences.

Our main contributions in this thesis are as follows:

- A new way of political preference extraction of Twitter users by applying image context extraction on Twitter profile images (Chapter 4),

- A method for finding profiles of people in Facebook given their name, surname and location (Chapter 5),

- A novel multi-stage computational framework for identifying users relevant to a topic by using existing communities in social media (Chapter 6).

1.2 Outline

Following is the structure of the thesis: In Chapter 2, we give preliminaries on the methods we used for identifying user interests in social media. Afterwards, in Chapter 3, we introduce related literature work. In Chapter 4, we describe a method for inferring supporters of political parties using Twitter profile images, and show the applicability of the proposed system through empirical data. In Chapter 5, we give the details of a basic user targeting method on Facebook, and display the obtained results. In Chapter 5, we propose a novel method for generalizing interest identification of social media users. In Chapter 5, we conclude this thesis by summarizing our contributions. In addition, we also propose some future research directions in identification of user interests in social media.

Chapter 2

Preliminaries

Before delving into the details of our work in user interest identification in social media, we present a brief introduction to methods we employ in order to extract interest of social media users from the multimedia content shared in public profiles. We apply natural language processing(NLP) methods on textual data; namely part-of-speech(POS) tagging, semantic similarity scoring and sentiment analysis. POS tagging helps us to detect core words in sentences such as nouns, semantic similarity scoring provides us a measurable distance between meanings of different words, and sentiment analysis extracts sentiments of users towards "things". We also consider images of social media users for identifying their interests. We apply image context extraction method for obtaining brief textual explanations of user images. These methods are described in length for a better understanding in the following subsections.

2.1 Part-of-Speech Tagging

A part of speech(POS) is a category of words defined by the morphological or syntactic nature of word being referred to. A list of POS tags are given in Table 2.1.1.

The assignment of allotting to each word in a textual content a part-of-speech like

noun or adjective is called part-of-speech tagging. POS tagging was first investigated in mid sixties and seventies [2, 3, 4]. Researchers engineered rules for this purpose back then.

Eventhough this task is easy for a computer to do when a language framed by rules and has a finite set of vocabulary, difficulties arise when it comes to dealing with POS tagging of natural language. Natural language refers to the language that humans use to communicate. It contains *unknown words*, *indeterminacies* in the sense that when there is disagreement about which tag is the correct one among human experts, and *noise* which means syntactic and morphological errors that are intuitive to humans to correct while communicating. Resolving such ambiguities is a hot-topic for POS tagging research.

Table 2.1.1: A practical list of POS tags in English

POS Tags	*Meaning*	*Examples*
ADJ	adjective	new, good, high
ADV	adverb	really, already
CNJ	conjunction	and, or, but
DET	determiner	the, a, some
EX	existential	there, there's
FW	foreign word	esprit, quo
MOD	modal verb	will, can, would
N	noun	year, home
NP	proper noun	Africa, April
NUM	number	one, 2014
PRO	pronoun	he, their, her
P	preposition	on, of, at, with
TO	the word to	to
UH	interjection	ah, bang, ha
V	verb	is, has, get, do
VD	past tense	said, took, told
VG	present participle	making, going
VN	past participle	given, taken
WH	wh determiner	who, when

Consider the following sentence as for applying POS tagging:

We speak not only to tell other people what we think, but to tell ourselves what we think. Speech is a part of thought.[1]

[1]Oliver Sacks, Seeing Voices

Given this sentence to a part-of-speech tagger, a part-of-speech is assigned to each word. In table 2.1.2, we can see that every word has a tag attached.

Table 2.1.2: Example of a tagged text

Word	*POS-Tag*	*Word*	*POS-Tag*
We	P	tell	V
speak	V	ourselves	N
not	ADV	what	PRO
only	ADV	we	P
to	TO	think	V
tell	V	.	<PUNC>
other	ADJ	Speech	N
people	N	is	V
what	WH	a	DET
we	N	part	N
think	V	of	P
,	<PUNC>	thought	N
but	CNJ	.	<PUNC>
to	TO		

In this thesis work, we use NLTK POS tagging software for our tagging purposes [5].

2.2 Semantic Similarity Scoring

Words are more similar when they share more features of meaning whereas they are less similar in case they have fewer common meaning elements. Measuring distance between *senses* of two words is called semantic similarity scoring.

Hyponymy-Hypernomy relationship is one of the most basic relationship between

senses of words. In linguistics, hyponomy is a word in a set of words grouped by meaning that refers to a specific subject. A hyponymy is included within the set referring to another word called hypernym. Computer science often refers this relationship an "IS-A" relationship descibed in Figure 2.2.1. For example, “automobile”, “bus”, “jeep”, “limousine”, “pickup”, “truck” and “van” are all hyponyms of hypernym “vehicle”.

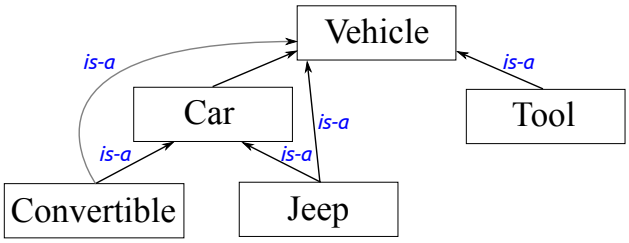

Figure 2.2.1: Hyponymy (IS-A) relationship

This taxonomy idea is feasible to be applied for making similarity judgements on noun pairs. There are alternatives to relating words in an IS-A taxonomy such as a wheel’s being a part of a car, snow’s being made up of water etc. Such non–hierarchical relations (e.g. has–part, is–made–of, is–an–attribute–of etc.) can be described additionally. Once neighbours of a word is organised, measurable path length between words can be used for word similarity measurement. Elaborations on this idea provided different measurement functions defined over distances. Some examples of such distance metrics are *lch* [6], *wup* [7] and *path similarity*.

Figure 2.2.2 displays a simple example on defining neighborhood and path distance between two words. For example, path distance between “rim” and “tool” is 4.

These relations are implemented in WordNet::Similarity software package [8]. We use this package for our similarity measurement purposes.

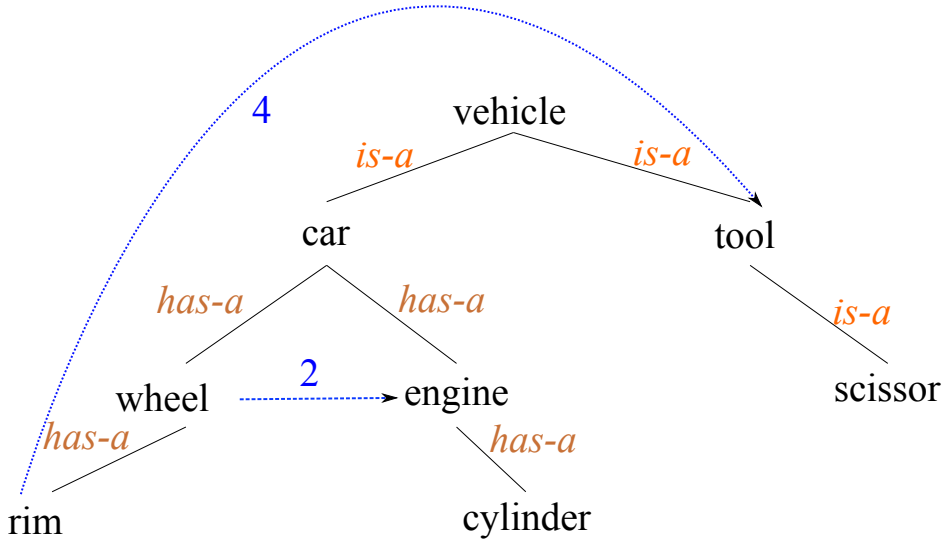

Figure 2.2.2: Path length on word relation graph

2.3 Web Image Context Extraction

Image Context Extraction is a method for bridging the semantic gap in text-based and content-based image representation methods [9]. While images are mostly accompanied by textual data in web (e.g. full-text, metadata, tags, figure captions etc.) extracting qualitative descriptions about these images requires methods that are capable of finding out the image context using accompanying textual data to each image. However, textual information may fail due to unannotated images or web pages in different languages. Therefore, depending only one web-source for this extraction can lead wrong results. For improving this situation by reducing the noise in the textual data, one way is increasing this textual data by searching for the same or similar images online. Considering and comparing different resources increases the chances of correct context extraction. In order to reach a set of same/similar images, online reverse image search can be used.

Reverse image search is a content-based image retrieval (CBIR) querying method that provides the CBIR system with an image which will be considered as a query. CBIR system retrieves this image, constructs a representation containing the information about

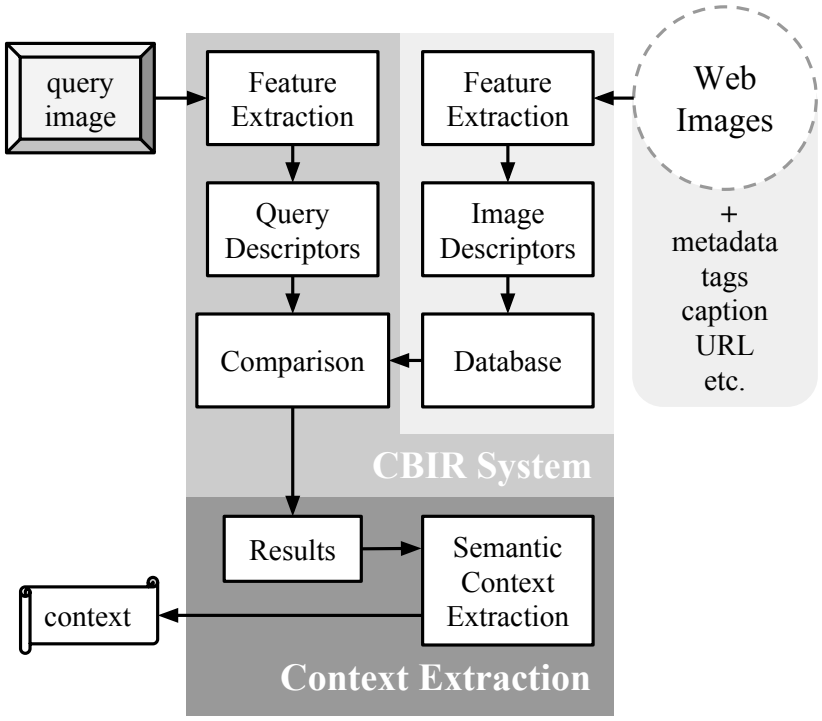

Figure 2.3.1: Web image context extraction (simplified)

its content, compares it with representations(i.e. Perceptual hashing [10]) of other images in a database and returns a related set of images [11]. There are two basic approaches in CBIR:

Low-level (local) feature approach: Retrieval by representing low-level features of images such as color, texture and shape.

Semantic approach: Retrieval by representing images as collection of objects and their relations.

Applying CBIR in web provides all metadata and tags of images along with the text in the page where the image is displayed. All this data can be used for inferring a context for the query image. Eventually, a *best-guess* context is obtained. Figure 2.3.1 simply describes image context extraction process.

In our work, we used Google Images [12] for image context extraction.

2.4 Sentiment Analysis

Textual data can be roughly categorized into two classes: facts and opinions. While facts are objective expressions, opinions are mostly subjective. They usually describe people's sentiments and feelings towards entities. Following are examples of factual and opinionated expressions [13]:

Factual statement

> *"The ocean is the connected body of salty water that covers over 70 percent of the Earth's surface."*

Opinionated statement

> *"Irresponsible oil companies have devastating effects on oceans."*

Eventhough understanding others' opinions are very important not only for individuals but also for organizations in decision making, the literature on processing opinions has recently developed. One of the major reasons of this is that, opinionated text became widely available after the World Wide Web. Before the web, people needed to ask for others' opinions before making a decision and organizations were conducting opinion pools when they needed to understand their customers or the sentiment of the public about their products. By the exponential growth of user-generated content on the web, this procedure is reformed. Nowadays, if a consumer is curious about a product, he/she can go online and search for opinions of other consumers. This invention allowed organizations to follow the opinions of their customers in a timely-fashion.

All these have become possible by the development of automated opinion discovery and summarization systems. *Sentiment analysis*, also known as opinion mining, is a result of this need. Due to its value and practical applications, both research in academia and applications in the industry have focused on improving this method. However, it is a challenging task due to natural language processing. Subtasks of sentiment analysis can be broadly categorized as follows:

Objectivity vs subjectivity: It is about determining whether a sentence expresses an opinion or not.

Opinion orientation: It is commonly known as sentiment classification. This process aims to discover the sentiment of the author in an opinionated text. For example, given a product review, it finds out whether the user has positive, negative or neutral feelings about the product.

Object extraction: Opinions can be expressed on different things (e.g. A product, a service, an individual, an organization, an event, or a topic). Object extraction finds these "things" on which the opinion is targetted.

Imagine analyzing the following sentence:

"Engines and gearboxes of BMW Z series are superlative."

This sentence is a *subjective* sentence. Opinion orientation is *positive* while its on *BMW Z series*.

In this thesis work, we use an implementation of sentiment classification procedure detailed in [14, 15, 16] for sentiment analysis purposes.

Chapter 3

Related Work

User interest identification has been a hot-topic for a long time. It helps social media service providers to recommend users personalized content, to infer their preferences and to categorize them respectively. Major focus of this domain is analysing user generated content. In this chapter we discuss the related work about targeting user interests.

3.1 Text Analysis and Semantic Analysis

Text analysis is usually described as a sub-field of data mining aiming at deriving high-quality information from textual data [17]. It involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, sentiment analysis, visualization, and predictive analytics [18]. The goal is, mainly, to obtain data from text by applying natural language processing and analytical methods. As user interest identification depends on examining textual data of users, methods in text analysis are very crucial for our purposes.

In [19], authors analyze social media graph, textual user interactions and group memberships to infer user interests and fields of expertise of users. They propose a technique

to obtain a similarity between a posts of users and their groups. In [20], a frequency-based method is employed for keyword extraction in microblogs that identifies interests of users accurately and efficiently. In [21], a search and summarization framework extracts relevant representative tweets of users in Twitter to generate a coherent and concise summary of different events.

Semantic analysis offers explicit ways to specify "things" related to users by developing standard measurable word sense relations. Therefore, needs, preferences, and activities of users are connected in a way that different systems can benefit from each other by sharing data data. This offers a wide range of benefits in the knowledge extraction and increased the results accuracy of user interest identification systems [22].

In [23], Wikipedia is used as a knowledge base for categorizing extracted terms from tweets. Categories of these terms are discovered for user profiling and interest extraction. A similar study [24], again make use of Wikipedia for ontological mapping of user interests for user clustering. Moreover, they integrate WordNet into their framework that leads better word sense disambiguation in extracted terms. In [25], they exploit semantic concepts captured from informal content of users such as messages and posts to build a semantic network which reflects people expertise.

3.2 Sentiment Analysis for Social Web

Sentiment analysis plays an important role in identifying user interests. As interests has a direction towards objects, activities, locations; more generally, towards "things"; sentiment analysis helps us to understand the sentiments of users on such "things". More specifically, sentiment analysis deals with the computational treatment of sentiments and opinions expressed in texts [26].

In [27], ontological resources are combined with natural language processing techniques to calculate polarity degree of user interests. In [28], semantic lexicons are created in order to identify sentiment words in blogs and news corpora. Then, a polarity value is

attached to each word in the lexicon. Such polarity is revised when a modifier appears in the text. Results shows the interests of users in these blogs and online news resources. In [29], topic-related opinionated texts are discovered and a sentiment score is calculated for each user post. In [30] they extract objects from tweets and added its semantic concept as an additional feature, and measure the correlation of the representative concept with negative/positive sentiment.

3.3 Recommender Systems for Social Media

Nowadays, social media analysis and social mining are used in recommender systems. These systems represent user interests on items. Effective recommendation includes filtering methods to predict and to suggest items that pleases users. These systems act as personalized decision guides, aiding users in decisions on matters related to personal taste. The key concept to develop an efficient recommender system is developing a better understanding of both users and interest items.

In individual user recommendation, the usual goal is to retrieve interest items with the highest scores, also referred to as relevance rating, computed by a strategy investigating users relation to an item. In [31], they identify topics and entities (e.g. persons, events, products) mentioned in tweets for a personalized news recommendation system. They consider temporal dynamics of those profiles, therefore quality of recommendations are improved by collecting more data over time. In [32], relationship information among people, tags, and items, is collected and aggregated across different sources for making recommendations related to user's topics of interest.

In group recommendation, however, an item may have different relevance to different group members and this disagreement among members must be resolved. In [33], authors analyze group data and propose a formal semantics that accounts for both item relevance to a group and disagreements among group members. In [34], they propose a group recommendation method that utilizes both social interests and item interests of group

members.

The latest trends in recommender systems domain takes into account how human beings function with their peers, especially in their interpersonal behaviors, which brings it closer to the field of social media analysis. In [35], they show how to infer interests for new users and inactive users from social media. Thus, they proposed an approach to combine text information and link information of users (information about social connection) to infer interests for inactive users. In the other study [36], they examine the range of data modalities such as text, network links, and categorical labels. They focused on a scalable machine learning system to visualize and explore the interests of millions of users on Facebook. Therefore, they obtain the relationship between concepts, user text and friendships.

3.4 Community detection

Community detection in social media is a tool for detection and analysis of characteristic communities. Detected communities often forms around topics of interests. It enables us to monitor opinions and interests, and provides a valuable insight about trends. Most of these opinion patterns are globally scarce yet locally dense in social media [37]. Social media and graph theory researchers are involved with inferring structures by examining these patterns among users and web pages.

In [37] all cliques in a social media graph are enumerated to find out maximal cliques of communities and these communities are adjusted by merging fractional communities. Therefore each community is labelled by a set of words describing interests of these communities. In [38], network, utility, feature and partition integration schemes are applied on different network dimensions of social media graph. Communities are extracted and labelled. In [39], the problem of community detection in the context of Social Media is explained on a wide variety of algorithms and their methodological principles.

In our work, we combine techniques form different categories and introduce novel

approaches incorporating context extraction from multimedia data. In Chapter 4, we use an method based on image context extraction technique. Textual data obtained from image context extraction is represented in a multiset of unique words. The more frequent a term appears, the more interested a user assumed to be in this term defining the topic. In Chapter 6, we introduce a method takes advantage of semantic analysis methods. We apply semantic analysis on textual data and multimedia data collected from users to obtain a score indicating the level of their interests in a predefined topic.

Chapter 4

Inferring Supporters of Political Parties in Twitter

This work examines supporter ratios of two mainstream political parties (Justice and Development Party(AKP) and Republican People's Party(CHP)) using profile images of Twitter users collected just before the 2014 local elections held on 30/05/2014 in Turkey.

Twitter is a popular micro-blogging service that allows messages of up to 140 characters to be posted and received by its users. These messages are called tweets. Tweets form the basis of interactions in Twitter where a user is updated about the tweets of other people he/she is following. Considering that there are more than 30 million active user of Facebook [40], and more than 9 million in Twitter [41] in Turkey, social media is an effective tool for analysing political trends[1].

Analyzing user content in Twitter gives an insight on public opinion in different topics. This content is also used as an indicator of political tendencies in many parts of the world [43, 44, 45]. As media stations and newspapers are known to have some degree of political bias, liberal, conservative or other; a considerable amount of work is done in providing objective results from Twitter data [46, 47, 48, 49, 50, 51]. These works em-

[1]Population estimation is 76,667,864 in 2013 [42]

phasize the usefulness of receiving diverse opinions shared in social media and conclude that inferring political atmosphere from Twitter provides healthier results than following traditional media resources.

There are many different ways for analysing public opinion in tweets. One simple approach would be defining a set of terms about the topic and counting the frequencies of these terms in collected tweets. This method gives an idea about which term is spoken more than the other. However, the opinion on these terms is not determined in this way. In order to determine in which tone (i.e. positive, negative) these terms are used in tweets, sentiment analysis is useful.

Our technique does not involve in text analysis. Instead, we collect profile pictures of users who post tweets including a set of terms related to these political parties. We assume that users presenting themselves by pictures having a "context" about these terms are positively associated with these parties. Image context extraction method is employed for obtaining textual information from these images. For this purpose, we use the Google Images [12].

4.1 Categorizing Twitter Users via Profile Pictures

For collecting texts, we defined two set of terms related to AKP and CHP. They are presented in Table 4.1.1.

Table 4.1.1: Term sets for two political parties

Term Set for AKP	Term Set for CHP
AKP	CHP
Recep Tayyip Erdoğan	Kemal Kılıçdaroğlu
Recep Tayyip	Mustafa Sarıgül
RTE	Atatürk
Osmanlı	Mustafa Kemal
Rabia	Gezi

4.1.1 Dataset and Method

We collected 100,000 tweets using BirdWatch application [52]. BirdWatch connects to the Twitter Streaming API and receives all Tweets that include at least one of terms in our term set defined for describing two different political tendencies. From collected tweets we extracted 65,000 unique twitter users including their usernames, names and profile picture URLs. This list of users L_U is our dataset.

For each user in L_U, we query Google Images with his profile picture URL and obtain either a *best-guess* context or an empty result. These results are collected in a frequency dictionary.

The URL of user image may become invalid in some cases. For this reason, we first check if the URL is valid. In case it is invalid, we try to extract user's profile image URL from his profile page by querying Twitter with his user name. However, users can also change their user names and querying twitter with a non-existing username returns an error page. If this happens, as a last chance, we try to find the user's new username via Google Web Search Engine [53] by a textual query including his name and "twitter". Supposing that we obtain user's profile page, then we can obtaim a valid profile image URL. Otherwise, we skip this user. We continue with this operation until each user is

processed. A flowchart describing the mechanism is presented in Figure 4.1.1.

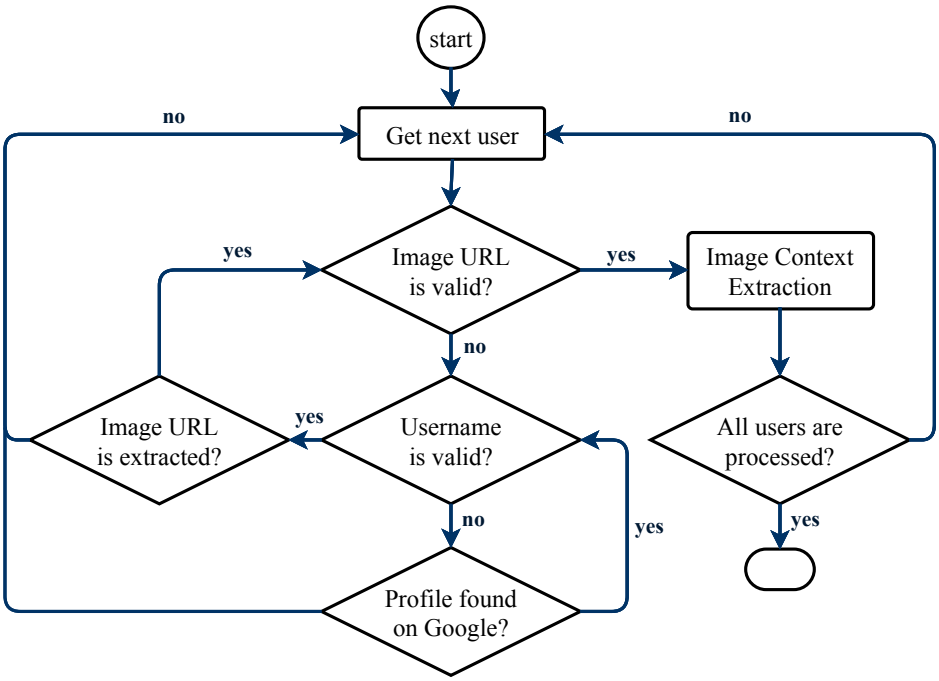

Figure 4.1.1: Flowchart of the system

Our framework is implemented in Python programming language.

4.1.2 Results

We manage to retrieve best-guess context results for 50,174 users out of 65,000(77%). 1,274 of these results are including one of our terms. Results are displayed in Figure 4.1.2.

Table 4.1.2: Results

Term Set for AKP		Term Set for CHP	
AKP	4	CHP	19
Recep Tayyip Erdoğan	27	Kemal Kılıçdaroğlu	4
Recep Tayyip	53	Mustafa Sarıgül	11
RTE	197	Atatürk	235
Osmanlı	118	Mustafa Kemal	89
Rabia	382	Gezi	135
Total	781		493
%	61		39

4.2 Post-Election Evaluation

In the local elections of 2014, the main party AKP received 42.87% of the votes while the second party CHP received 26.34% of the votes [54]. When we disregard the votes other parties received, proportions are 62% for AKP and 38% for CHP. These numbers are very similar (61% and 39%, Table 4.1.2) to our results. This result shows the real potential of our work in inferring supporter proportions of political parties.

Chapter 5

Finding Audience for Advertisement

Social media allows users to communicate, share knowledge about similar interests, discuss favorite topics, review and rate products/services, etc. It accommodates a big potential for businesses if it is correctly integrated in the marketing mix [55]. The basic marketing principles are still valid, yet, companies must now be creative in order to target audiences and make a profit.

Social media marketing can be very profitable for businesses. If approached correctly, it can help to find talent, to build brand awareness, to target new customers, and to conduct brand intelligence and market research [56]. For example, viral propagation via friends communicating among each other and creating user engagement by building brand applications are two possible ways of marketing strategies in social media [57]. Moreover, social marketing is cost-saving. It is an inexpensive way to promote a business rather than hiring a marketing team [58]. The advertising problem problem is essentially a problem of matching. As J. Wanamaker, a pioneer in marketing once said, "Half the money I spend on advertising is wasted. The trouble is, I don't know which half". In social media, automated use of keywords and content targets the users that are most likely relevant to the advertised product. Considering the number of users in social media, the advertisement reaches a large and *correct* audience.

5.1 Searching Facebook Profiles via Name and Places Lived

Facebook is a popular free social networking website that allows registered users to create profiles, upload photos and video, send messages and keep in touch with friends, family and colleagues. In our work, we find Facebook profiles of people whose names and locations are provided in a list. The results will be used for advertisement purposes.

5.1.1 Dataset and Method

We are given a list of 1500 people from a particular city in Turkey. The process starts by querying the Facebook Graph API [59] to receive a list of matches for each entry in the given list. For some common Turkish names, the Facebook Graph API returned up to 800 possible results. To refine such list of profiles for each name, we visited and checked "Places lived" section of each profile in a parallel fashion. In case a location match occurs, the found profile is associated with the name. A flowchart describing the mechanism is presented in Figure 5.1.1.

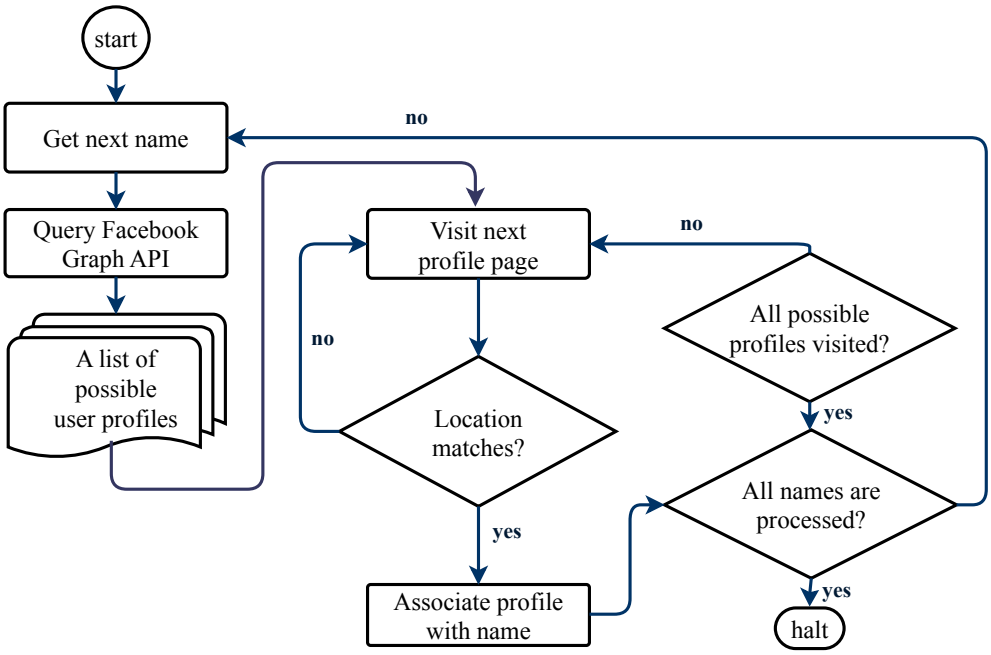

Figure 5.1.1: Flowchart of the system

5.1.2 Results

1,332 out of 1,500 names (88%) are matched to Facebook profiles whose "Places lived" section are matched. This result shows the applicability of automated targeting of Facebook profiles via names and locations.

Chapter 6

Relevance-based Condensation of Online Communities in Social Media

In this chapter, we describe our a novel framework for creating a community of a particular interest. For this purpose, we analyze the users in existing communities about this interest, and select the users that are defined as related by our framework. We name this process *community condensation*.

6.1 Introduction

The most important characteristic of the humans is tendency to join with others in communities. In sociology, an offline community (often termed as a group) is defined as a number of people who communicate and interact with each other. People join these communities for many reasons. Even the notions of communities are complex, multidimensional, and in flux; an intuitive approach would suggest that it satisfies needs of people that are either difficult or impossible to access individually. They also benefit from numerous default elements of being a member in such a community. Some of them are namely as follows: [60]

Companionship – Communities provide company of other people.

Survival and security – History and evolution shows that hunting and defense requires humans to act together.

Affiliation and status – Some members in different communities can obtain certain social statuses or securities.

Power and control – Communities bring opportunities for leadership roles for individuals who are strong in their opinions and can exert it over others.

Achievement – Overlapping of interest and motivation creates success.

By the internet revolution, as restrictions of location and time in communication methods teared down [61], boundaries become more permeable and connectivity ever more possible. As a consequence, individual, social, organizational, professional and political conceptions and configurations are undergoing a vast change. Not long ago, technological foundations of Web 2.0 gave birth to a group of Internet-based applications called *Social Media* [62]. This unprecedented invention has led to the proliferation of online communities by gathering people around mutual interests, helping to share knowledge and to spread ideas quickly and providing support among community members in a timely fashion [63]. An example of online community formation in social media is shown in Figure 6.1.1.

These progresses contested the meaning of "offline" communities. Due to accessibility and efficiency of online communities, they are being considered offering an equivalent value in community formation. As a supporting evidence, primary needs for participating in online communities remained very similar of those previously listed for offline communities [64, 65].

Undeniable potential of online communities may not be always fulfilled and many communities fail. Communities heavily depend on individuals to develop and maintain them. Overlaps on personal interests and shared focus among members is a must-have for productive relationships in communities. Unfortunately, most of the communities suffer from heterogeneous interests and purposes of their members due to random member accessions. This provokes untidiness in communal interactions and ends up with lack of

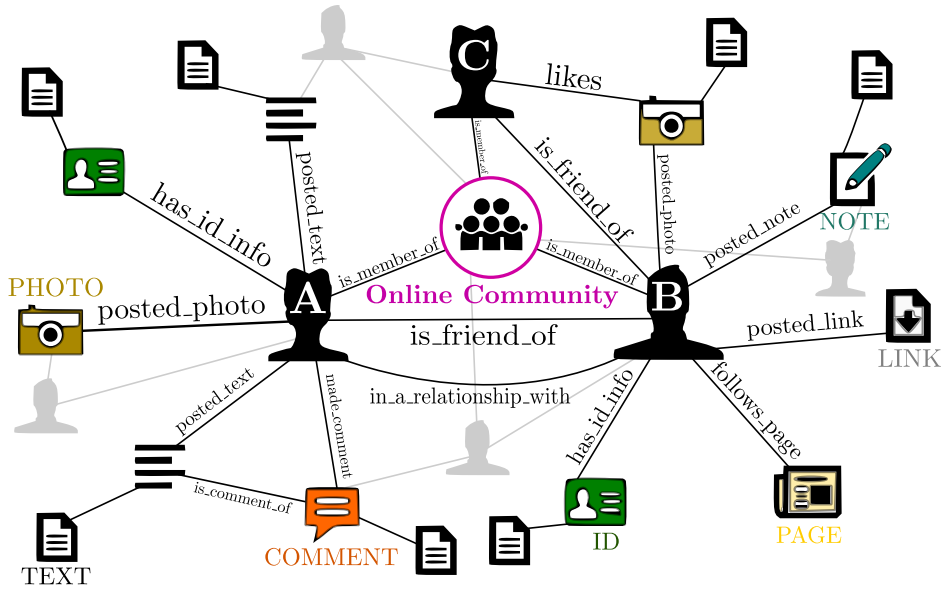

Figure 6.1.1: Online Community in Social Media

interest and under-contribution in the community [66].

We attack this problem by targeting highly-interested users on a given topic to create a *core* community. Our novel multi-stage computational method integrates various techniques adopted from information retrieval, natural language processing and computer vision for estimating a quantitative relevance score for each user indicating his/her interest to the topic. Afterwards, we rank users and form a *condensed* community from the indivudials having a high score. A straightforward approach would be scoring all users in a social media. However, social media service providers do not share a complete list of their users, and even if they do, computing a relevance score for all users would be computationally infeasible. An alternative feasible solution is condensing existing communities about the targeted topic into a smaller community consisting of users whose scores are above a threshold score. An overview of our approach is given in Figure 6.1.2.

Bridging highly-related users together in a condensed community helps these people to serve in the same direction and to build a productive community. Our framework can be highly effective in bringing people together around personal interests and hobbies; gath-

ering activists around issues; creating discussion and support communities; and targeting consumers for advertisement.

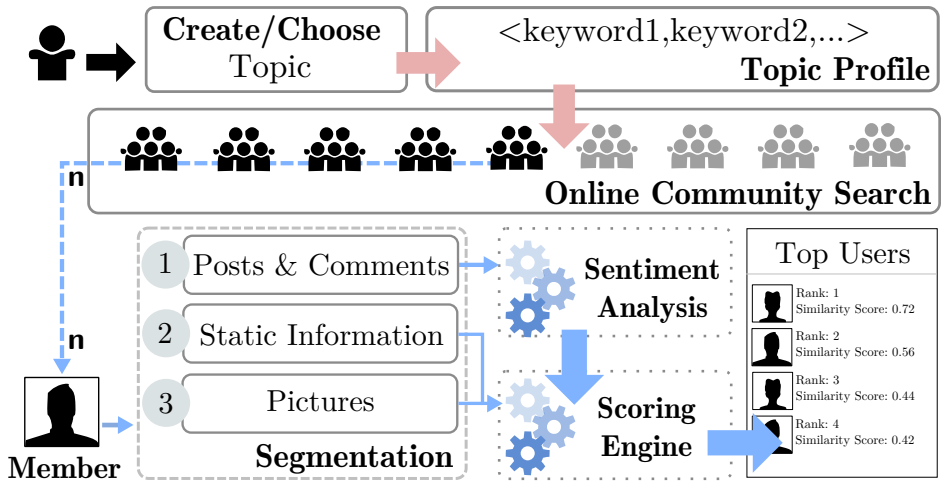

Figure 6.1.2: Overall Architecture

This chapter is organized as follows: First, we define the relevance-based condensation problem in Section 6.2. The proposed framework is introduced in Section 6.3. In Section 6.4, we show the applicability of our framework on Facebook Groups and exhibited results of a user survey to demonstrate the correctness of our framework. Finally, a discussion is presented in Section 6.5.

6.2 Relevance-based Condensation of Communities

A perfect relevance scoring function $\gamma(p_u)$ simulating human cognitive processes to label a user either as relevant or irrelevant to an interest topic T by going through the public information on each user's profile p_u for all users U in a social media SM, would provide us a set of related users U_γ to T. However, such a simulating function does not exist and even if it does, applying a function on all users in a social media would be computationally infeasible. Moreover, social media service providers do not reveal full list of users U.

We propose to start from an initial seed set of users rather than all users. In order

to find this initial set of users, we take advantage of online communities that are already present and formed around certain interests in SM. As shown in Figure 6.2.1, our hypothesis is that sets of users of communities about topic T intersects with U_γ. If this hypothesis holds, our proposed method should target some of these users in intersection.

We first search for communities C_T in SM using the set of keywords defining the interest topic in K_T. Afterwards we extract their members U_C who are possibly interested in T. Our relevance scoring function $r(p_u)$ estimates a score for each user $u \in U_C$ using p_u. By selecting users having a score above a threshold t, we target a set of users ${U_T}^t$ that are interested in T. We call this process condensation of existing communities. The higher the threshold value t is, the more relevant users are accumulated in ${U_T}^t$. We experimentally verified this claim in Subsection 6.4.3. Note that, it is not possible to say that a user is irrelevant if his/her score is below t since relevance score depends on self-disclosure[1] of users. Our hypothesis about intersection of U_C and U_γ is also justified by asking experiment participants to label our results obtained by $r(p_u)$ in Subsection 6.4.3.

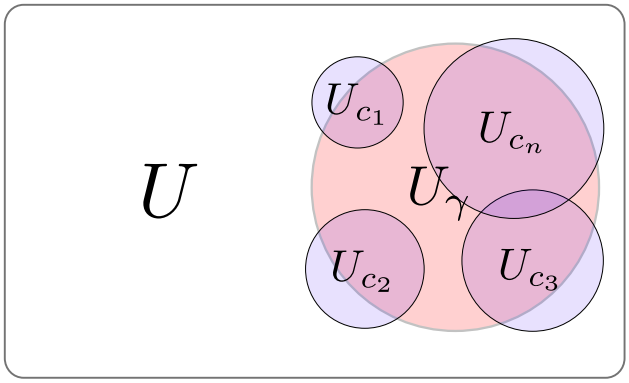

Figure 6.2.1: Role of communities in reaching highly-relevant users

Now, we can introduce the details of our framework.

[1]Self-disclosure is defined as what individuals publicly reveal about themselves to others, including thoughts, feelings, and experiences [67]

6.3 An Integrated Semantic Approach to Relevance-based Condensation of Communities

We propose a novel framework for solving relevance-based condensation of communities, that consists of three stages as depicted in Figure 6.3.1. These stages piece together some techniques of information retrieval, NLP and image context extraction in the following subsections.

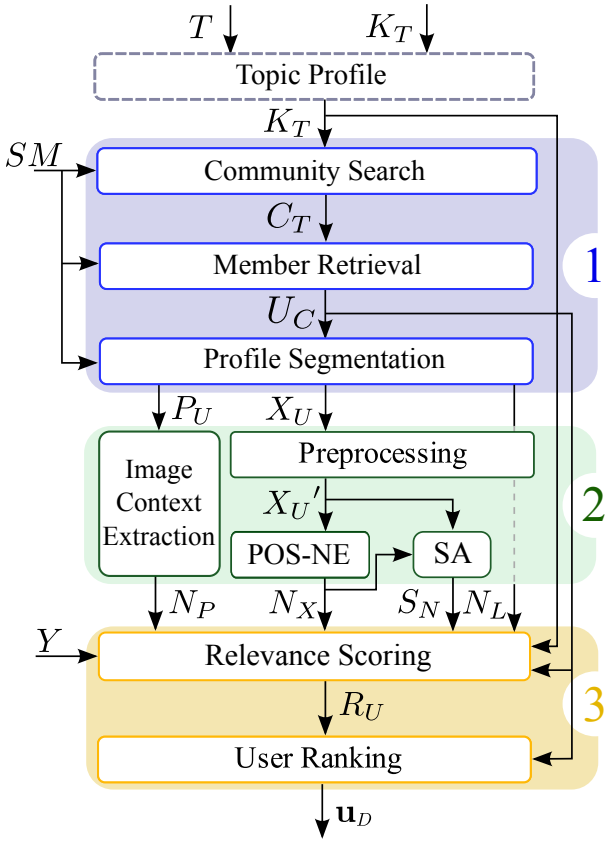

Figure 6.3.1: Framework Flowchart: (1) Information Retrieval, (2) Information Extraction, (3) Scoring Engine

6.3.1 Information Retrieval

First, we search in a social media SM for a set of online communities C_T using the keyword set K_T for topic T. After that, we collect all members of communities in C_T in a set of users U_C. As our aim is targetting highly relevant users in U_C, we need to collect what individuals reveal about themselves on their public profiles in SM.

Posts are the most common way for users to express their thoughts and emotions in their profiles. We gather this textual data for each user in a set X_U. Along with textual data, pictures provide essential clues about lifestyles and interests. Therefore, public pictures of each user is collected in P_U. In some social networking sites there is a particular section for displaying things users care about (e.g. Likes in Facebook and +1s in Google+). We create another set N_L for storing such data in textual format for each user. While collecting this information, users who do not share any public information in their profiles are assumed to prefer staying private and are naturally excluded from the set U_C.

We are now ready for extracting information about collected users described in the following subsection.

6.3.2 Information Extraction

Given retrieved textual data X_U and images P_U, our target is to extract nouns from X_U by applying NLP methods and textual context from P_U using image context extraction.

Identifying nouns provides us key words that determine the meaning of sentences. For example, if we are trying to condense online communities in a domain related to politics, the posts of their members would likely contain names of some politicians and political parties. This implies that nouns are a set of relevant terms we target.

Before applying term extraction process on the set of user posts X_U, we preprocess this textual data in two steps. First, we ensure that the language of elements in X_U are in English. The language of the sentences are detected using NLTK language classifier [5] and translated in English if necessary. For this translation task, we use online machine-

translation service Google Translate [68, 69]. Second, sentences are normalized to help Part-of-Speech(POS) tagging for noun extraction(NE). This process replaces abbreviations on auxilary verbs (e.g. "I'm" $\rightarrow$ "I am"), and smileys (e.g. ":-)"$\rightarrow$ "happy"). Once this process is completed, preprocessed set of sentences ${X_U}'$ is ready for POS tagging.

We employ NLTK POS Tagger for identifying nouns [5] in sentences. It marks each word in a text to its corresponding part of speech. For example, "A beautiful day!" is respectively tagged as determiner, adjective, noun and punctuation mark. Given preprocessed set of sentences ${X_U}'$, nouns are collected in a set of terms N_X. At the same time, we apply an unsupervised sentiment analysis method(SA) detailed in [15, 16] on ${X_U}'$ for obtaining polarity values of sentences. This set of polarity values S_N help us to determine the sentiment of users towards extracted nouns.

While textual information of users are being processed, we query online reverse image search engine Google Images [12] to obtain a textual context in a few words describing each image in P_U. These results are stored in N_P. Since the elements of the set N_P and N_L are brief titles and descriptions, we treat these texts as isolated nouns and directly use them for relevance scoring.

6.3.3 Relevance Scoring

In this subsection, we calculate relevance score of each user in U_C by measuring the semantic similarity between the keywords in K_T and extracted nouns for each user in N_X, N_P and N_L that are obtained in the previous stage. For estimating a relevance score between two words, we use WordNet Path Similarity [8] measure.

This quantitative relation between senses of words are based on information contained in a thesaurus hierarchy graph. For example, a "road" might be considered more like a "path" than a "car", if "path" has a common ancestor with "road". Another possibility is that "path" has a closer ancestor to "road" than "car" in hierarchy graph. Intuitively, the shorter the path length between the words is in graph, the more similar their senses are.

Algorithm 1: Relevance scoring

Input: U_C, a set of users,
K_T, a set of domain keywords,
N_X, a set of terms from posts of users,
N_P, a set of terms from images of users,
S_N, sentiment analysis values of N_X,
N_L, a set of terms from likes of users,
Y, a set of scoring constants for pictures and likes.

Output: R_U, relevance scores of users.

```
// S_N = {s : 0≤s≤1}
R_U = {}
foreach user u in U_C do
    // Score of the user is 0 at the beginning
    score_u ← 0
    N_Xu ← get terms of user u from N_X
    foreach term t in N_Xu do
        M = {}
        s_n ← get SA value of noun n from S_N
        foreach keyword k in K_T do
            M.append(sim_path(n,k)*s_n)
        score_u += getMaxValue(M)
    foreach noun n in N_P do
        if n in K_T then
            score_u += getPictureConstant(Y)
        else
    foreach noun n in N_L do
        if n in K_T then
            score_u += getLikeConstant(Y)
        else
    R_U.append(score_u)
return R_U
```

The definition of WordNet Path Similarity metric is as follows:

$$pathlen(w_1, w_2) = \#\text{edges on the shortest path}$$

$$sim_{path}(w_1, w_2) = \text{-log}pathlen(w_1, w_2)$$

For each user u, the highest path similarity scores between extracted nouns from his posts N_X and keywords in K_T are added up on his/her score $score_u$. Therefore, we infer a number indicating a degree of relevance from each user's posts. Afterwards, the nouns obtained from pictures N_P and likes N_L are searched for a string match in K_T. In case of matchings, predefined constants are added up on $score_u$. The whole relevance scoring process is detailed in Algorithm 1.

The final step is ranking users in U with respect to their relevance scores R_U. The ranked user vector $\mathbf{u}_T$ is the output of our framework containing the users in an order of relevance to domain T.

We apply our method on Facebook Groups in the following section.

6.4 A Use-Case Scenario: Implementation and Experiments with Facebook Groups

6.4.1 Experiment Setup and Dataset

We selected a topic T about *Cars* because transportation is a common need in societies and cars are among popular topics in social media. Our first step is creating a list of topic keywords as follows:

$$K_D = \{\text{Cars, Automobile, BMW, AUDI, Mercedes, Volkswagen, Ford, Renault, Toyota, Honda}\}$$

Figure 6.4.1 exhibits number of members in groups obtained by group search in Facebook by using each keyword in K_T. We disregarded private groups and groups that have

size smaller than 50. In total, 557 public groups collected in set of groups C_T and about 1.1 million users are retrieved in set of users U_C. This suggests that we can reach approximately 100000 users per keyword.

Due to data collection limitations of Facebook, we randomly selected 2500 users from this set of users U_C for profile segmentation and applied our condensation methods on this subset U'_C of U_C.

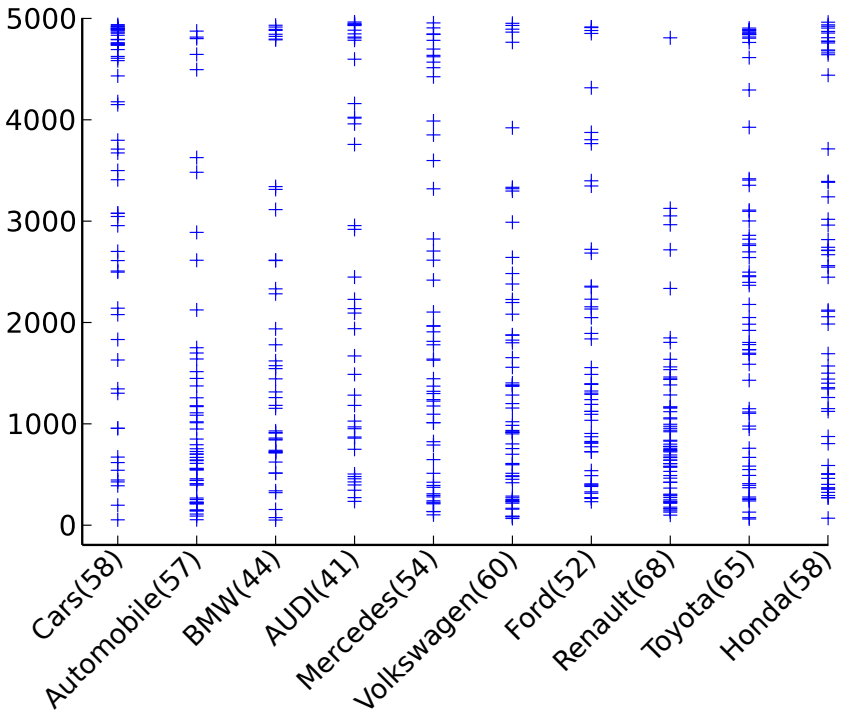

Figure 6.4.1: Number of groups obtained by each keyword and group sizes

6.4.2 Results

During profile segmentation of users in U'_C, we found out that about every 2 users out of 3 have either very limited publicly available content or none. Therefore, scores of 800

users out of 2500 are calculated and normalized by scaling between 0 and 100 as shown in Figure 6.4.2.

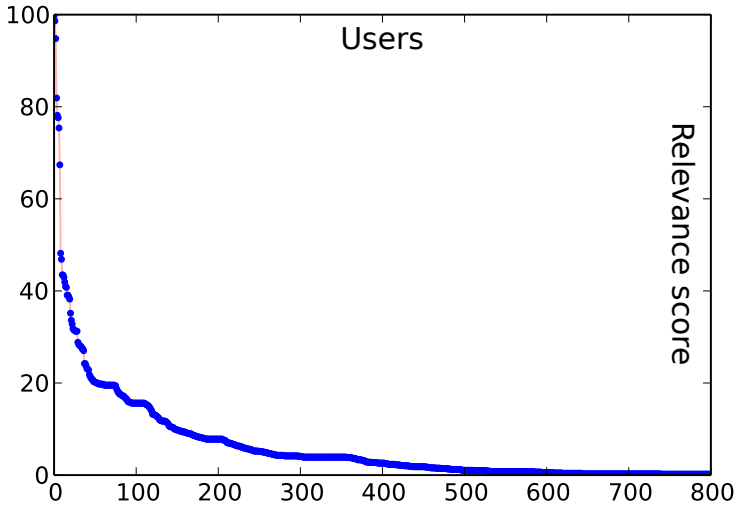

Figure 6.4.2: Relevance scores of U'_C.

When a threshold t were defined over the results obtained, the number of users proposed for the condensed group would be inversely proportional to t as it is displayed in Figure 6.4.3.

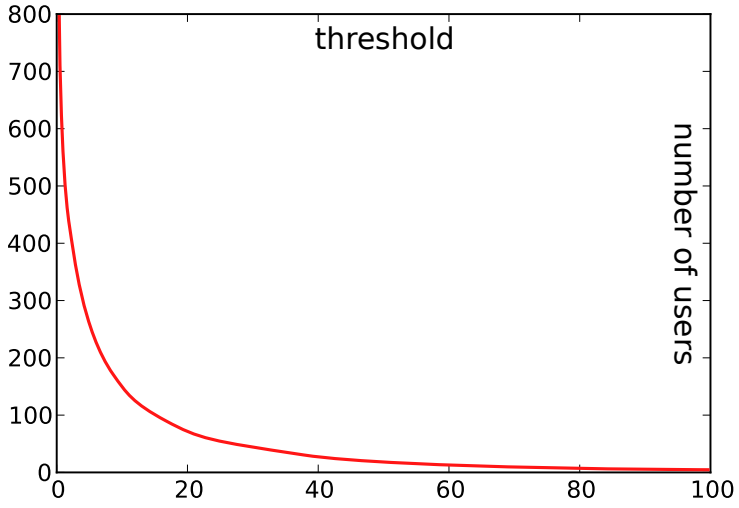

Figure 6.4.3: Threshold vs number of users

6.4.3 Validation

We demonstrate the effectiveness of our proposed computational method via a user study. First, we defined score intervals for labelling users in three categories. These categories are namely *Highly-relevant*(HR), *Relevant*(R) and *Irrelevant*(IR) users(See Figure 6.4.2). Afterwards, we asked three people to label 100 randomly chosen users from U'_C. They labeled these users by going through available public data in their Facebook profile pages.

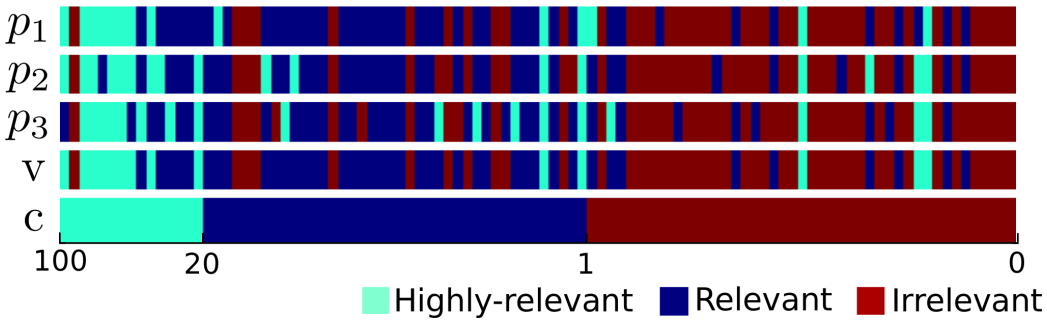

Figure 6.4.4: Validation result arrays

Results of user labellings are exhibited as color coded label arrays of users in Figure 6.4.4.

Array u_1, u_2 and u_3 represent labellings of three experiment participants. We aggregate these three results by voting three answers for each profile. This result is displayed on array v. In array v, we obtained 14 highly-relevant, 41 relevant and 45 irrelevant labels. For deciding on score thresholds of these three labels, we first sort the relevance scores of 100 users, and then defined the thresholds on 14th and 55th users. Therefore, threshold scores of labels are decided as follows:

- Highly-relevant users ($relevanceScore \geqslant 20$)
- Relevant users ($20 >$relevanceScore$\geqslant 1$)
- Irrelevant users ($relevanceScore < 1$)

According to this labeling scheme, we labelled the results of our framework on 100 users. This labeling is shown in array c in Figure 6.4.4.

Table 6.4.1: Validation results using 3 labels

	HR	*R*	*IR*	Overall 1
u_1	53	67	71	66
u_2	60	60	71	64
u_3	53	57	73	62
v	60	70	73	70

Table 6.4.1 and Table 6.4.2 exhibits matching percentages of user labellings to framework results. In Table 6.4.1, we make the comparison using three labels as it is previously described. We achieved 70% of matching in overall.

Table 6.4.2: Validation results using 2 labels

	HR+R	*IR*	Overall 2
u_1	80	71	76
u_2	75	71	73
u_3	77	73	75
v	80	70	76

In Table 6.4.2, we consider *Highly-relevant* and *Relevant* labels as one label shown as *HR+R*. By doing this we inspect the differentiation capability of our framework between *Relevant* and *Irrelevant* users. In overall, we achieved 76% of matching to user labellings. This result shows that our hypothesis in Section 6.2 about the intersection of U_C and U_γ is correct.

We verify the functionality of threshold in community condensation process by using the users labelled by experiment participants. As threshold value gets higher, we expect

to obtain communities formed of more relevant users. For justifying this statement, we calculated percentages of relevant users in communities using the labelled group members by increasing t. Assume that U_t is the set of users obtained by using threshold t and U_r is the set of users labelled as related by participants in U_t, percentage of relevant users at a threshold t calculated as $|U_r|/|C_t|$. In Figures 6.4.5,6.4.6, and 6.4.7, we clearly see that as threshold value is set higher, the proportion of relevant users increase in the condensed community. Thus, our expectations about setting a threshold in Section 6.2 are justified.

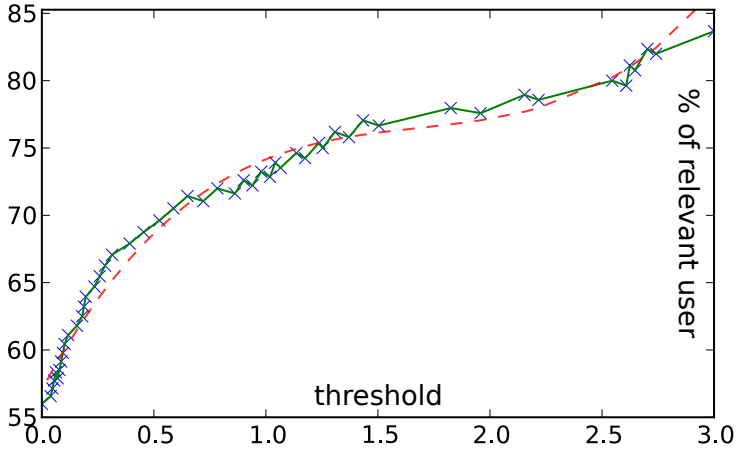

Figure 6.4.5: Threshold (0-3) vs target percentage

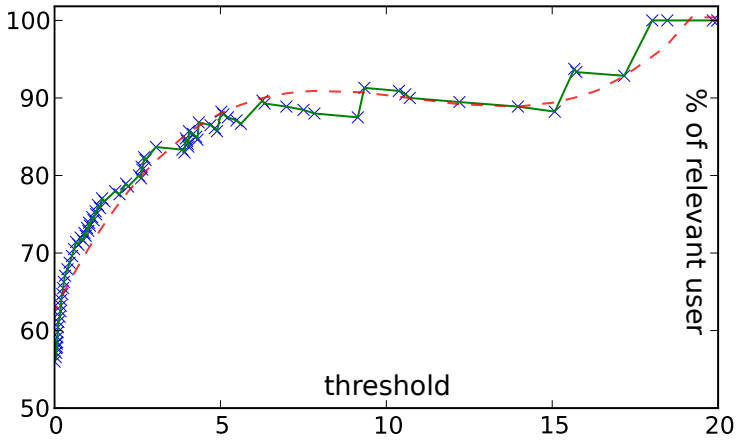

Figure 6.4.6: Threshold (0-20) vs target percentage

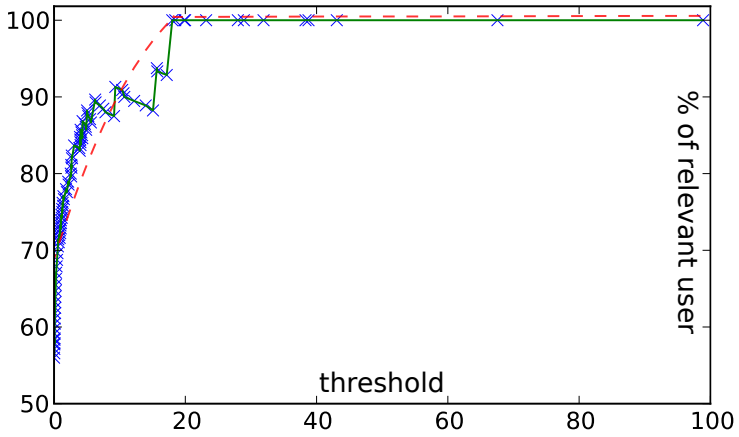

Figure 6.4.7: Threshold (0-100) vs target percentage

6.5 Discussions

The obtained results indicates the potential of our framework that can increase the value of current online communities by bringing hobbyists, enthusiasts, activists, consumers and people who share a common curiosity together. This work also affirms a solution to problems of today's online communities such as under-contribution.

We plan to improve our framework by using domain ontology that conceptualize user relevance to a particular topic or area of interest. We hope this initial work serves as a motivation to research seeking solutions to enhance online communities.

Chapter 7

Conclusions

In this thesis, we have proposed three methods for identifying users and their interests from multimedia data shared by users. In Chapter 4, we used image context extraction on profile pictures of Turkish Twitter users for inferring proportions of two different political stances before March 2014 local elections. The percentages we obtained using our system just before the elections were only 1% different from the proportions in election results. This result suggests the variety of Turkish Twitter users and it emphasizes the effectiveness of social media analysis in forecasting political tendencies in Turkey.

In Chapter 5, we found Facebook profiles of people whose identification information (i.e. Name and location) is provided. We obtained 88% of the initial list including identification information of targeted people. This high rate matching rate demonstrates the usefulness of our approach.

In Chapter 6, we presented a multi-stage computational framework for condensing communities that are related to a topic in social media. We collected users of communities that are already present about a given topic and applied a relevance-scoring technique through the public information these users share on their profile pages. Our results show that our method accurately differentiate between users who are interested in given topic and those who are not interested in it. We illustrated the functionality of defining a thresh-

old score for obtaining communities in different sizes and concentrations. We validated the results of our framework by a user-study. We found 76% of match between user labelled and automated results. This promising result suggests that our framework that can increase the value of online communities.

In the future, we want to integrate our user interest identification systems to domain ontologies which are structural frameworks for organizing information in semantic web. This direction will allow us to understand user-user and user-interest relations better. We will also be able to apply rule based reasoning methods on these ontologies for inferring new knowledge from observable user data.

Bibliography

[1] ComScore. *It's a social world: working and where its headed. Top 10 need-to-knows about social net*. URL: http://www.comscore.com/Insights/Presentations-and-Whitepapers/2011/it_is_a_social_world_top_10_need-to-knows_about_social_networking (visited on 06/2014).

[2] Zellig Harris. "String Analysis of Language Structure". In: *Mouton and Co., The Hague* (1962).

[3] Sheldon Klein and Robert F Simmons. "A computational approach to grammatical coding of English words". In: *Journal of the ACM (JACM)* 10.3 (1963), pp. 334–347.

[4] Rubin G. M. Greene B. B. "Automatic grammatical tagging of English. Technical Report". In: *Brown University, Department of Linguistics* (1971).

[5] Steven Bird. "NLTK: The Natural Language Toolkit". In: *Proceedings of the COLING/ACL on Interactive Presentation Sessions*. COLING-ACL '06. Association for Computational Linguistics, 2006, pp. 69–72.

[6] Claudia Leacock and Martin Chodorow. "Combining local context and WordNet similarity for word sense identification". In: *WordNet: An electronic lexical database* 49.2 (1998), pp. 265–283.

[7] Zhibiao Wu and Martha Palmer. “Verbs semantics and lexical selection”. In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 1994, pp. 133–138.

[8] 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. Boston, Massachusetts: Association for Computational Linguistics, 2004, pp. 38–41.

[9] Alberto Del Bimbo. *Visual Information Retrieval*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. ISBN: 1-55860-624-6.

[10] Christoph Zauner. “Implementation and Benchmarking of Perceptual Image Hash Functions”. MA thesis. Austria: Upper Austria University of Applied Sciences, 2010.

[11] B. Szanto et al. “Sketch4match x2014; Content-based image retrieval system using sketches”. In: *Applied Machine Intelligence and Informatics (SAMI), 2011 IEEE 9th International Symposium on*. 2011, pp. 183–188. DOI: 10.1109/SAMI.2011.5738872.

[12] Google Inc. *Google Images*. URL: http://images.google.com/imghp (visited on 06/2014).

[13] Bing Liu. “Sentiment analysis and subjectivity”. In: *Handbook of natural language processing* 2 (2010), pp. 627–666.

[14] Nikolaos Pappas. *unsupervised$_s$entiment*. URL: https://github.com/nik0spapp/unsupervised_sentiment (visited on 01/2014).

[15] Nikolaos Pappas, Georgios Katsimpras, and Efstathios Stamatatos. “Distinguishing the Popularity between Topics: A System for Up-to-Date Opinion Retrieval and Mining in the Web”. In: *Computational Linguistics and Intelligent Text Processing*. Vol. 7817. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 197–209.

[16] Nikolaos Pappas and Andrei Popescu-Belis. “Sentiment Analysis of User Comments for One-class Collaborative Filtering over Ted Talks”. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’13. ACM, 2013, pp. 773–776.

[17] J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006.

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

[19] Akram Alkouz, Ernesto William De Luca, and Sahin Albayrak. “Latent Semantic Social Graph Model for Expert Discovery in Facebook.” In: *IICS*. Ed. by Gerald Eichler et al. Vol. P-186. LNI. GI, 2011, pp. 128–138.

[20] Zhiyuan Liu, Xinxiong Chen, and Maosong Sun. “Mining the interests of Chinese microbloggers via keyword extraction”. In: *Frontiers of Computer Science* 6.1 (2012), pp. 76–87.

[21] Freddy Chong Tat Chua and Sitaram Asur. “Automatic Summarization of Events from Social Media.” In: *ICWSM*. 2013.

[22] R. Studer, R. Benjamins, and D. Fensel. “Knowledge engineering: principles and methods”. In: *Data and knowledge engineering* 25 (1998), pp. 161–197.

[23] Matthew Michelson and Sofus A. Macskassy. “Discovering Users’ Topics of Interest on Twitter: A First Look”. In: *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*. AND ’10. Toronto, ON, Canada: ACM, 2010, pp. 73–80. ISBN: 978-1-4503-0376-7.

[24] M. Wasim et al. “Extracting and modeling user interests based on social media”. In: *Multitopic Conference (INMIC), 2011 IEEE 14th International*. 2011, pp. 284–289.

[25] Anna Lisa Gentile et al. “Extracting semantic user networks from informal communication exchanges”. In: *The Semantic Web–ISWC 2011*. Springer, 2011, pp. 209–224.

[26] Hsinchun Chen and David Zimbra. “AI and Opinion Mining.” In: *IEEE Intelligent Systems* 25.3 (), pp. 74–80.

[27] Juana María Ruiz-Martínez, Rafael Valencia-García, and Francisco García-Sánchez. “Semantic-Based Sentiment analysis in financial news”. In: *International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012)*. 2012, p. 38.

[28] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. “Large-Scale Sentiment Analysis for News and Blogs”. In: *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*. 2007.

[29] Nikolaos Pappas, Georgios Katsimpras, and Efstathios Stamatatos. “Distinguishing the Popularity between Topics: A System for Up-to-Date Opinion Retrieval and Mining in the Web”. In: *Computational Linguistics and In-*

telligent Text Processing. Vol. 7817. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 197–209.

[30] Hassan Saif, Yulan He, and Harith Alani. "Semantic Sentiment Analysis of Twitter". In: *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*. ISWC'12. Springer-Verlag, 2012, pp. 508–524.

[31] Fabian Abel et al. "Analyzing user modeling on twitter for personalized news recommendations". In: *User Modeling, Adaption and Personalization*. Springer, 2011, pp. 1–12.

[32] Ido Guy et al. "Social media recommendation based on people and tags". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2010, pp. 194–201.

[33] Sihem Amer-Yahia et al. "Group Recommendation: Semantics and Efficiency." In: *PVLDB* 2.1 (Sept. 2, 2009), pp. 754–765.

[34] A. Beach Q. Lv M Gartrell X. Xing and R. Han. "Enhancing Group Recommendation by Incorporating Social Relationship Interactions". In: *Proceedings of the 2010 international ACM conference on Supporting group work (Group 2010)*. 2010.

[35] Tingting Wang et al. "Mining User Interests from Information Sharing Behaviors in Social Media". In: *Advances in Knowledge Discovery and Data Mining*. Vol. 7819. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 85–98.

[36] Qirong Ho et al. "Understanding the Interaction between Interests, Conversations and Friendships in Facebook". In: *CoRR* abs/1211.0028 (2012).

[37] Nan Du et al. “Community detection in large-scale social networks”. In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM. 2007, pp. 16–25.

[38] Lei Tang, Xufei Wang, and Huan Liu. “Community detection via heterogeneous interaction analysis”. In: *Data Mining and Knowledge Discovery* 25.1 (2012), pp. 1–33.

[39] Symeon Papadopoulos et al. “Community detection in social media”. In: *Data Mining and Knowledge Discovery* 24.3 (2012), pp. 515–554.

[40] Socialbakers Staff Writer. *Turkey is Facebook world country No. 4*. URL: http://www.socialbakers.com/blog/207-turkey-is-facebook-world-country-no-4 (visited on 06/2014).

[41] Ahmet Can Sit. *Comparison of Turkey and Brazil digital industries by numbers*. URL: http://en.webrazzi.com/2012/09/17/comparison-of-turkey-and-brazil-digital-industries/ (visited on 06/2014).

[42] Wikipedia. *Turkey*. URL: http://en.wikipedia.org/wiki/Turkey (visited on 06/2014).

[43] Axel Bruns and Tim Highfield. “Political networks on twitter: tweeting the Queensland state election”. In: *Information, Communication & Society* 16.5 (2013), pp. 667–691.

[44] Alex Burns and Ben Eltham. “Twitter free Iran: An evaluation of Twitter’s role in public diplomacy and information operations in Iran’s 2009 election crisis”. In: (2009).

[45] Anders Olof Larsson and Hallvard Moe. “Studying political microblogging: Twitter users in the 2010 Swedish election campaign”. In: *New Media & Society* 14.5 (2012), pp. 729–747.

[46] Jennifer Golbeck and Derek Hansen. “Computing political preference among twitter followers”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2011, pp. 1105–1108.

[47] Jisun An et al. “Media Landscape in Twitter: A World of New Conventions and Political Diversity.” In: *ICWSM*. 2011.

[48] Adam Bermingham and Alan F Smeaton. “On using Twitter to monitor political sentiment and predict election results”. In: (2011).

[49] Marko Skoric et al. “Tweets and votes: A study of the 2011 singapore general election”. In: *System Science (HICSS), 2012 45th Hawaii International Conference on*. IEEE. 2012, pp. 2583–2591.

[50] Erik Tjong Kim Sang and Johan Bos. “Predicting the 2011 dutch senate election results with twitter”. In: *Proceedings of the Workshop on Semantic Analysis in Social Media*. Association for Computational Linguistics. 2012, pp. 53–60.

[51] Hao Wang et al. “A system for real-time twitter sentiment analysis of 2012 us presidential election cycle”. In: *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics. 2012, pp. 115–120.

[52] Matthias Nehlsen. *BirdWatch*. URL: `https://github.com/matthiasn/birdwatch` (visited on 06/2014).

[53] Google Inc. *Google*. URL: `http://www.google.com/` (visited on 06/2014).

[54] Anadolu Ajansı. *2014 Yerel Seçim Sonuçları*. URL: http://secim.haberler.com/2014/ (visited on 06/2014).

[55] Victoria Bolotaeva and Teuta Cata. "Marketing opportunities with social networks". In: *Journal of Internet Social Networking and Virtual Communities* 2010 (2010), pp. 1–8.

[56] Rustey Weston. *7 Social Networking Strategies*. URL: http://www.entrepreneur.com/article/191312 (visited on 06/2014).

[57] Christy Pettey. *Gartner Says Social Networks Are Attracting Too Much Traffic for Retailers to Ignore*. URL: http://www.gartner.com/newsroom/id/660409 (visited on 06/2014).

[58] Christy Pettey. *5 Reasons Why Social Marketing is a Must. Relativity Business Technology Solutions*. URL: http://www.relativitycorp.com/socialnetworkmarketing/article1.html (visited on 06/2014).

[59] Facebook. *Introducing Facebook Graph Search*. URL: https://www.facebook.com/about/graphsearch (visited on 01/2014).

[60] Steve M. Jex and Thomas W. Britt. *Organizational Psychology: A Scientist-Practitioner Approach*. Wiley, 2008. ISBN: 9780470196472.

[61] Jenny Preece. "Sociability and usability in online communities: determining and measuring success". In: *Behavior and Information Technology* 20 (2001), pp. 347–356.

[62] Andreas M. Kaplan and Michael Haenlein. "Users of the world, unite! The challenges and opportunities of Social Media". In: *Business Horizons* 53.1 (2010), pp. 59 –68.

[63] Joon Koh and Y-G Kim. “Knowledge sharing in virtual communities: an e-business perspective”. In: *Expert Systems with Applications* 26.2 (2004), pp. 155–166.

[64] Andreas Girgensohn and Alison Lee. “Making Web Sites Be Places for Social Interaction”. In: *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*. CSCW ’02. New Orleans, Louisiana, USA: ACM, 2002, pp. 136–145. ISBN: 1-58113-560-2.

[65] Namsu Park, Kerk F Kee, and Sebastián Valenzuela. “Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes”. In: *CyberPsychology & Behavior* 12.6 (2009), pp. 729–733.

[66] Pamela J. Ludford et al. “Think Different: Increasing Online Community Participation Using Uniqueness and Group Dissimilarity”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’04. Vienna, Austria: ACM, 2004, pp. 631–638. ISBN: 1-58113-702-8.

[67] Kathryn Greene, Valerian J Derlega, and Alicia Mathews. “Self-disclosure in personal relationships”. In: *The Cambridge handbook of personal relationships* (2006), pp. 409–427.

[68] Python for Facebook. *facebook-sdk*. URL: `https://github.com/pythonforfacebook/facebook-sdk` (visited on 01/2014).

[69] Google Inc. *Google Translate*. URL: `http://translate.google.com` (visited on 01/2014).