Babes-Bolyai University Cluj-Napoca
Faculty of Economics and Business Administration

Research Field: Cybernetics and Statistics

On Effective Methods for Social Media Content Analysis and Recommendation

PhD Thesis Summary

PhD Supervisor:
Prof. Dr. Tomai Nicolae

PhD Candidate:
Cristina Ioana Muntean

-2012-

Thesis Contents

Abstract

The expansion of social networks and large amount of data from Social Media give rise to interesting applications and technologies that support them. We start our research by first trying to understand Social Media and the mechanisms that engage users to interact on these platforms. As a case study we select Twitter on which we will base all our observations, study and applications. We study it from a double perspective, that of the user and that of social network analysis. We wish to approach the processing of data collected from Twitter with the adequate tools, thus we inspect Big Data and MapReduce technologies for processing large amount of data. We study the models for text representation and recommendation by exploring insights from several disciplines: Information Retrieval, Machine Learning and Recommendation Systems. After looking for related work to ours regarding Twitter hashtag clustering and recommendation, we propose our own models. We extract a large dataset from Twitter for a period of three weeks, representing 10% of the public stream. We clean the data and preprocess it. On the resulting dataset we make a thorough analysis in order to find insights and understand the specificity of the problem. We notice that often hashtags have a cryptic sense (acronyms, concatenation of words and numbers). We wish to explore the meaning of hashtags by clustering them together into groups. We see that top terms in clusters succeed in translating their meaning giving an overall view on the context in which they appear. We also propose a hashtag recommendation system. After looking at hashtagging patterns we discover their dual nature, the first one is replacing regular words or phrases in a tweet with a hashtag consisting of that word or phrase (inline hashtags), and the second one is offering the context, an informal category to which that tweet can be attributed considering its semantics (context hashtags). We propose a model that encompasses both behaviors by treating the recommendation of inline hashtags as a prediction problem and the recommendation of context hashtags as an information retrieval problem. The proposed system outperforms the state of the art, but also propose solutions to how results can be further improved.

Keywords: Social Media, Gamification, Big Data, Twitter, Hashtag clustering, Hashtag recommendations

Contents

Chapter 1

Introduction

Users as well as companies on the Internet need to understand the power at their disposal and take advantage of the opportunities that might arise from the large amount information coming their way. The fact that the use of Internet has exploded in the last couple of years, puts us in the position of having to filter, sort and select information that is of use to us, from a sea of unstructured, unreliable and huge quantity of data. This may not be as transparent for simple Internet users, but for example Internet companies such as Google, Twitter, Facebook, Amazon, Last.fm, deal with this, and more, in their services. They strive to optimize data storing, transform information into knowledge and offer relevant personalized information, products and services to their users, in attempt to maximize their satisfaction.

Our goal is to study how we can improve and stimulate the interaction between users and available tools, platforms and services provided over the Internet.

1.1 Motivation

The fact that Internet users benefit from a lot of facilities nowadays, enables us to discover different methods to build and further improve them. Our motivation lies in discovering new ways to use and exploit present information, overwhelming in its complexity and size, in order to ease the use and create novel tools which can be of help to various actors activating online. Internet data is evolving and changing rapidly and research, in its respect, must keep the pace.

Due to the interdisciplinary approach, in our research we employ various research methods, thus a multi-method research, which take into account several dimensions of the problem. At the same time we offer pragmatic solutions

for the set objectives, i.e. short text analysis, clustering and recommendation applications for a particular Social Media environment, Twitter.

1.2 Objectives

Due to the fact that Social Media has emerged recently, we are confronted with the fast shifts and developments in this field. The challenges and problems deriving from Social Media have changed during time, but in our study we try to keep the pace with this rapidly advancing topic. The applications we propose take into consideration new and interesting tasks which exploit its particular structure and content.

The main objective of this work is to find efficient methods for Social Media analysis and propose useful applications to social media users, in order to make the best of the facilities and information available in this environment.

For approaching the subject we set the following specific *objectives*:

1. We wish to make a thorough analysis of Social Media and available social networking sites which support information exchange between existent connections.

2. We wish to better understand the appeal Social Media platforms offer to users, what are the advantages and disadvantages of becoming an active component in such online communities and how users can be stimulated in order to engage more in this kind of activity, namely information sharing in a social context.

3. We wish to study the particularities of the content that can be found in these environments (usually short text) and discover relevant applications that could be used to improve user experience and data analysis. We want to achieve this by doing the following:

   (a) Study the related work regarding various applications and analyses regarding Social Media and microblogging platforms (which have the specificity of enabling communications through short text).

   (b) Collect and extract the information available on Social Media platforms.

   (c) Use the proper tools for analysis and modeling the huge amount of data available in such environments.

   (d) Consider the adequate algorithms and methodologies, like supervised and unsupervised learning, for mining the available content.

(e) Propose our own applications and confront them with existing state-of-the-art.

1.3 Thesis overview

Keeping in mind the objectives we have previously set, we focus our attention on possible ways to approach those objectives and offer possible solutions. An overview of the thesis is presented in Figure 1.1.

Part of the research presented in this thesis was made during a 10 month internship at the High Performance Computing lab, ISTI CNR, Pisa, under the supervision of Fabrizio Silvestri and together with research teams mentioned in [12, 7, 6, 25].

In Chapter 2 we present the social context on the Internet and how it has evolved in the recent years. We describe Social Media, identify important Social Media channels and exemplify by choosing Twitter as its exponent. We analyze the platforms and the services available and explain how it can be used in order to fully exploit its capabilities. In Chapter 3 we describe what makes users activate on social platforms and analyze user experience and engagement, and how it can be stimulated. A way of doing this is through Gamification. We explain in detail what it means, the impact it has, and Twitter is gamified and propose a model for gamifying a social e-learning web application. In Chapter 4 we explain how to collect data from Internet in general and Twitter in particular and what are the technologies that handle large scale data processing and modeling in a distributed way, a requirement when dealing with such big amount of information. In Chapter 5 we look at possible ways to model and recommend textual data and describe the methodologies we have used and implemented further in our research. We also show how the results can be evaluated. In Chapter 6 we make a detailed study of on how short text from Twitter can be approached in research and present other available applications proposed by research. We also look at state-of-the-art approaches in clustering and making text recommendations in Twitter. In Chapter 7 we present the analysis of our Twitter dataset and how we approached clustering and recommendations of hashtags from Twitter and motivate their usefulness. In Chapter 8 we present our conclusions and future research work.

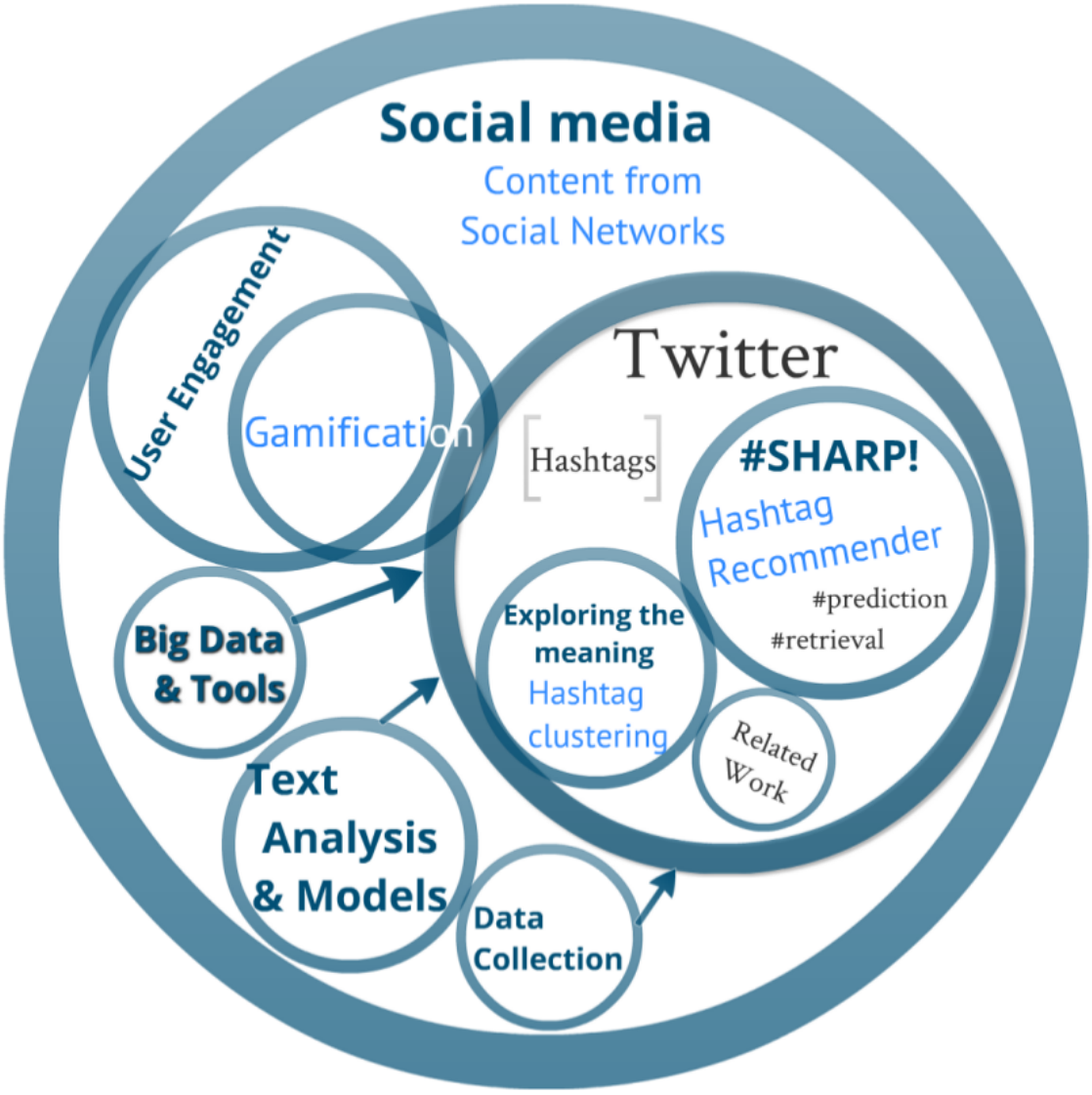

Figure 1.1: Thesis Overview

Chapter 2

Social media and Twitter

In this chapter we start by introducing essential concepts of the work. We briefly describe the interactions between users, customers and companies in the context of the social web and user generated content. User interaction means value creation for companies on the Internet and it is in their interest to stimulate user engagement and activity.

Tapscott and Williams [43], in their book "Wikinomics: How Mass Collaboration Changes Everything", elaborate on the new world of web-based economics where cultural values such as participation, collectivism and creativity are promoted. A few examples of such initiatives are YouTube, MySpace, Wikipedia, Flickr, Second Life, Linux or Twitter. They are created by crowds of anonymous users who express themselves in their own manner, with little constraints in comparison with the Web 1.0 model. Accordingly, instead of the hierarchical business model of producer–consumer, we can now encounter the so-called 'co-creation' model [41]. Through mass creativity, peer production and co-creation, the gap between collective (non-market, public) and commercial (market, private) production becomes smaller, as well as the distance between producers and consumers. We study the mechanisms behind Social Media: co-creation, marketing as conversation and open innovation and Social Media user behavior. According to Osterwalder [33, 34, 35], Social Media arises from a sum of platforms (blogs, wikis, social networks, micro-blogs and other platforms where users generate content) that allow and encourage social interactions through content sharing and discussions.

In [22] the authors state that Social Media represents a powerful force and that the executives who are reluctant or unable to understand and develop Social Media strategies, are missing out on opportunities while also being threatened by competitors which know how to take that advantage. The authors identify the main *functionalities* of Social Media: identity, conversations, sharing, presence, relationships, reputation and groups. Li and Bernoff

in [24] divide the Social Media participants according to their behavior in: creators, critics, collectors, joiners and spectators.

As already mentioned before, Social Media has many enablers, but the most part is supported by social networks. We will further direct our discussion towards Twitter, which we use as case study.

Twitter is a social network and microblogging platform with one way relationships between users that allows them to post short status messages, containing URLs, mentions and/or hashtags. These post can receive replace, be set as favorites or be retweeted. Twitter is also referred to as a real-time information network. We have analyzed Twitter from a double perspective, that of a user and from a social network analysis point of view, mentioning the most relevant studies involving Twitter as a social network.

A social network is the pattern of friendship, advice, communication or support which exists among the members of a social system [23, 11]. Twitter says about itself it is: "The fastest, simplest way to stay close to everything you care about." It binds a user to the latest stories, news, ideas and opinions, all the user has to do is to find the accounts he is interested in and follow them. We also offer an overview of the state-of-the-art in social network analysis regarding Twitter. It has been discovered that in spite of the one way relationships between users, the degree of separation between users is around four, users follow a power law and there is a very low degree of reciprocity between them. Users have different intentions: information sharing, information seeking and friend-wise relationships. The conclusion is that Twitter is not the typical social network, making it an interesting object of study.

Chapter 3

User engagement and Gamification

As previously mentioned, for companies activating on the web, interaction on behalf of the users (browsing, recommending, sharing and commenting) can increase their value. It is important to create a customer base and stimulate users to keep coming back to their website or share content with their social network. In this chapter we study user engagement, what are its characteristics and the metrics for measuring it.

According [3] engagement is "the emotional, cognitive and/or behavioral connection that exists, at any point in time and over time, between a user and a technological resource." The characteristics are: focused attention, positive affect, aesthetics, endurability, novelty, richness and control, reputation, trust and expectations, motivation, interests, incentives and benefits. In [38] the authors propose eight indices for measuring engagement: click depth index, duration index, regency index, loyalty index, brand index, feedback index and interaction index.

In the social context we must not only study the interaction of a user with a peace of content, we have to take into consideration also the social relationships and interactions between users.

Since we have previously stated its importance, we also propose a solution for stimulating user engagement, namely through Gamification. Gamification [17] is "the use of game-play mechanics for non-game applications". Any application, task, process or context can theoretically be gamified. Gamification's main goal is to rise the engagement of users by using game-like techniques such as scoreboards and personalized fast feedback [19] making people feel more ownership and purpose when engaging with tasks [36]. The main means of rising engagement through gamification are: accelerated feedback cycles, well defined rules and goals, compelling narrative and the tasks are

achievable. We explain the behavioral mechanisms behind engagement and gameplay. Game mechanics are characterized by following three attributes:

- Game mechanics type: Progression, Feedback, Behavioral;
- Benefits: engagement, loyalty, time spent, influence, fun, SEO, UGC, Virality;
- Personality types [8]: explorers, achievers, socializers and killers.

In this chapter we discuss various game mechanics and Gamification techniques and give an example for gamifying an application (using Points, Leaderboards, Status, Levels etc.). Twitter also is analyzed through the Gamification lens. By observing Twitter, we can find several *engagement statistics* and *feedback loops* that suggest Gamification: number of followers, retweets, direct messages, @-mentions, #-tags and lists. Even if there is no formal leaderboard throughout the platform, users do become aware of their status, by looking at all the elements above mentioned. Due to the fact they are public profiles, users become compelled, and enter the feedback loop of the service. Gamification is at the foundation of the service, even if less evident.

Chapter 4

Managing Big Data

In this chapter we take a more pragmatic look at the technologies used for collecting data from Social Media and other websites and what are the tools that can help process large quantities of data. We present two methods for collecting data: through web crawlers and through streaming API. We implemented these methods in order to collect data for two different applications.

We use a web crawler in order to crawl and collect data about eBay users and their corresponding feedback (positive, negative, neutral) from other users. Another crawler was created for extracting job posts for BestJobs. Both were implemented in Python.

For extracting data from Twitter we use the Twitter Streaming API. We were granted the Gardenhose access level, designed for research purposes offering 10% of the public stream (Sprizer, the standard accessibility account, offers only 1%). The average rate for download was at around 14.700 tweets per minute. We collected a large scale dataset of tweets over a period of three weeks consisting of 443,288,820 tweets overall, approximately 200 G of compressed files, the size increasing up to 10 times in an uncompressed format. The tweets are delivered in a JSON format.

In order to be able to store and process this amount of data, we explore technologies that deal with Big Data. According to Edd Dumbill of O'Reilly Radar [18], Big Data is the data that exceeds the processing capacity of conventional databases. The data is very dynamic, it is too large and does not fit the structures of databases.

In memory computations are have a high cost when it comes to processing such large amounts of data. According to [45] there are two determinant factors that have contributed to its development: the rise of massive social networking platforms and the development of Map Reduce [16]. Relational databases no longer fit this kind of data.

We present Apache Hadoop, Cascading and Apache Mahout, all frame-

works that allow parallel computing, which we have used for processing our data, as explained in Chapter 7. With the help of these frameworks and libraries we have been able to process data in MapReduce. Apache Hadoop offers an abstraction layer over a cluster. We were able to store the Twitter dataset on the Hadoop Distributed File System.

Cascading is an open source Java library, enabling MapReduce data processing over Hadoop clusters. From this data set we filtered out un-useful information and ran various Cascading jobs that cleaned the data from tweets in other languages other than English, tweets without hashtags, tweets with only URLs and hashtags (usually spam), etc. The input for any operation comes from a `Source`, it is processed into one or more `Pipes` that allow various functions or operations (Each, Every, GroupBy, CoGroup, SubAssembly) and later is delivered in output `Sinks`.

Apache Mahout is a MapReduce library that implements Machine Learning algorithms. It has three modules: Recommendations, Classification and Clustering. We expand the clustering section by speaking about the K-means implementation of Mahout, which we have used for clustering Twitter hashtags [31]. The basis for the development on this framework was set in [14], where authors explain how various learning algorithms can be parallelized.

Chapter 5

Models for data analysis & recommendations

After explaining what are the tools to use for handling large scale data, in this chapter we explore models for textual data analysis and recommendations. Because our objective was to analyze content from Twitter, cluster hashtags and offer hashtag recommendation, we have explained what are the models we have used in order to represent our tweets and hashtag collection. We have referred to both data structures and used algorithms.

The metaproblem at this part of the research is setting the basis for Twitter text analysis and recommendations. We tackle this problem from various angles, as the borders between methods are quite fuzzy.

We explain how our hashtag recommendation task can be approached from various perspectives: content based recommender systems, Information Retrieval and Machine Learning algorithms for classification or clustering. User may suffer from *interaction overload* and *information overload*. To overcome these problems several methods have been proposed in different but related fields such as *Information Retrieval*, *Information Filtering*, and *Recommender Systems*. We explain the basic principles of analyzing text, from feature selection, positional and conditional independence, vectors and weighing schemes (*tfidf*, *BM25*), to indexing documents and the inverted index. We also make an overview of Machine learning algorithms we have used in our experiments.

We have represented hashtags as "bag-of-words", any hashtags is composed of all the tweets in which that hashtag appears. We have selected the text features that are relevant and discard the others, like stop words, and we have unified related ones, by stemming. We have used the Vectors Space Model and $tfidf$ weighing to represent the body of each hashtag as vectors.

In the last section of the chapter we see what are the evaluation metrics we

will have to use for validating our results. We consider both ranked and non-ranked results evaluation metrics. We discuss precision, recall, F-measure, NDCG, P@k, MAP.

Chapter 6

Related work on Twitter applications

In Chapter 6 we have presented the related work and applications involving Twitter: information diffusion [49, 4], trend detection [26, 2], sentiment analysis [15, 21, 1], spam detection [44, 46, 9] or recommending real-time news [40, 42].

The discussion in the latter part of the chapter is strictly related to the applications we have proposed in Chapter 7. Having in mind our objective of clustering hashtags, we have narrowed down the discussion to clustering text and short text. Regarding the task of clustering short texts, in [5] authors cluster texts from RSS/Atom feed reader by enriching that scarce information from the feed with additional features from Wikipedia. Several papers on Machine Learning techniques applied to Twitter tackle subjects like summarization and topic detection (LDA) [32], clustering [48] and disambiguation of topics or classification [39, 37].

Most studies on clustering regarding Twitter include topic modeling algorithms. In [39] the authors use LDA in order to classify short and sparse text using hidden topics from large-scale data. Recommendation systems use clustering as a prior step to offering suggestions. In [13] the authors suggest tweets based on a user's history and topic model. They transform text according to vector space model and assign $tfidf$ weights to vectors. Similarly, TwitterRank [47] is based on tweet topics and the authors attempt to find influential users. They use LDA to build topic models for each users according to their tweets.

In [20], the authors analyze the use of hashtags for tweet tagging and compare tagging behavior of Twitter users and Delicious users. In [10] the authors try to identify important topics by taking into account the structure of the social network.

Lastly we have presented a similar work to ours, namely offering hashtag recommendations. To the best of our knowledge, [50] by Zangerle *et al.* is the only paper investigating hashtag recommendation. In this paper the authors exploit the similarity among the current and past tweets in order to generate hashtag recommendations based on previous activity of users. The most similar tweets to the one entered by the user are retrieved from a dataset of tweets by using $tfidf$ similarity. The hashtags from the resulting set of tweets are then extracted and ranked according to three different methods:

Overall Popularity Rank - ranks the hashtag candidates according to their popularity in the entire dataset.

Recommendation Popularity Rank - ranks the hashtags candidates according to the number of occurrences in the set of candidates.

Similarity Rank - calculates the similarity between two tweets by using $tfidf$, more precisely between the tweet currently entered and a tweet from the dataset, which provides the hashtag recommendation candidates.

Results are evaluated by means of precision and recall figures and show that *Similarity Rank* outperforms the other ranking metrics.

Chapter 7

Analyzing, clustering and recommending hashtags

In the last chapter we have presented the models, the results of the experimentation phase and their evaluation. If in Chapter 4 we have presented how data has been collected and selected, in the last chapter we have described the dataset obtained and some of the preprocessing steps we have taken. From the whole dataset we have eliminated tweets that are not in English, retweet, conversational tweet with mentions or spam tweets. We have presented also a couple of statistics regarding the dataset, we have noticed that the hashtags have a power law distribution, as can be seen in Figure 7.1, meaning few of them are very frequent while many of them occur just a few time. This is also due to the fact that hashtags are user generated content and there are no restrictions in creating them. They can include acronyms, spelling mistakes and so on; there is no hashtag validation step before they can appear in the network. The dataset is described in Figure 7.1.

Twitter Dataset – (December, 02 – December, 23)	
Days of activity	21
Tweets per day	21,108,991
Tweets per day (English only)	9,014,780
- with at least one hashtag	1,576,905
- without retweets and mentions	987,892
N. distinct hashtags per day	283,915

Table 7.1: Properties (on average) of the dataset we used for evaluation.

In order for the clustering and recommendation of hashtags to have sense, we first prove the utility of hashtags. Hashtags increase the visibility of a

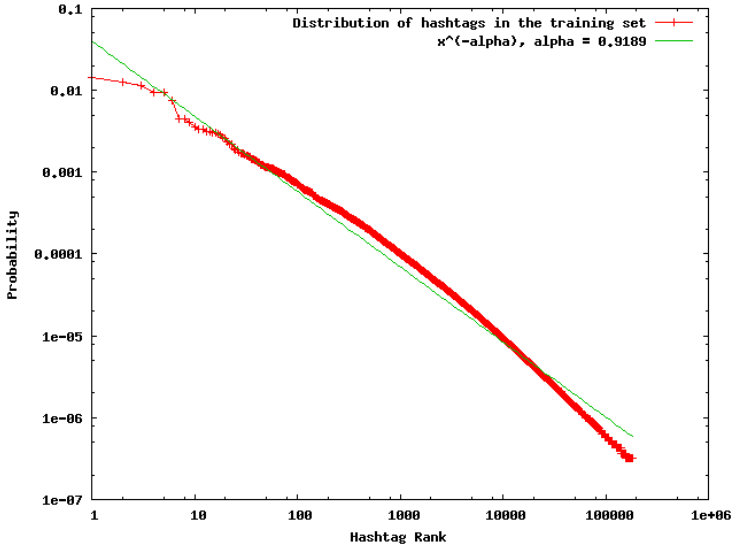

Figure 7.1: Probability distribution of the hashtags in the training set.

tweet above the friendship level, so by employing hashtags in a tweet, they can be visible in the entire network to those looking for that tweet or topic; hashtags can help group together and structure entire conversations. We see the probability of a tweet being retweeted[1] is significantly higher in cases where hashtags have been used, as can be seen in Table 7.2.

Probability	Value
$P(R\|H)$	0.25%
$P(R\|\overline{H})$	0.16%

Table 7.2: Conditional probability of retweet by varying the set of the given tweet t: i) containing hashtags $t \in H$, ii) not containing hashtags $t \in \overline{H}$.

The first application we describe is the clustering of hashtags, which has not been done before. The motivation behind it is rather simple, but this application can be used in other various contexts as well. As stated before we have noticed that hashtags have lot of particularities, they are made of acronyms, concatenated words, concatenated numbers and words, basically they cannot be interpreted in an automatic way and can be cryptic for users as well. Our idea is to cluster similar hashtags together based on the tweets in which they appear. For each hashtag we have created a so called virtual document, which works as a dictionary for the meaning of that term. We

[1]The retweet is a reinforcement mechanisms for the quality of a tweet.

want to capture their semantic meaning from similarity with other more explicit hashtags, also represented by their virtual document, or with top terms from the clustered documents, based on similarity. In these experiments we use daily datasets, which we have transformed in ¡hashtag, virtual document¿ associations and cluster the vectorized documents using K-means. We experiment with a variable number of clusters, k=20, 40, 80, 100, 200, ..., 1000. The results are encouraging. If for a small number of cluster the groups are not very clear, for a bigger number of clusters, k¿100, the clusters are better delimited and the topic can be easily identified. We have presented the top hashtags in a cluster along with top terms, the correspondence is obvious as can be seen in Table 7.3. We calculate the inter and intra-cluster distance, in Figure 7.2, and the results show that for k¡100, the clusters are not very well delimited, whereas for k¿100 they are better defined. We use k as a kind of precision or granularity of the topics/groups we want to discover. For a big k the topics are more precise, while for a smaller k the topics are more general.

top terms	occupy, ows, wall, street, protest, ndaa, movement, afghanistan, noccupy, st.
top hashtags	ndaa, ows, occupy, occupywallstreet, china, peace, yyc, economy, kpop, washington.

Table 7.3: Cluster example for Dataset15 with $k = 500$

The second application is *recommending hashtags*. For this task we build a system called #SHARP! , as described in Figure 7.3. We analyze hashtags and discover a logical division according to their functionality and purpose. Accordingly, hashtags are divided in inline hashtags (hashtags replacing words from a tweet) and contextual hashtags (hashtags that characterize the topic of the tweet and are not included in the tweet as terms). According to the two types of hashtags, we treat the recommendation problem from a dual perspective: we have treated the recommendation of inline hashtags as a prediction problem and the recommendation of contextual as a retrieval problem. We reunite the recommendation results in a single list, combined the results and presented the user with the final list of recommendations. We thoroughly describe how the recommendation is modeled and run the experiments as described.

The results have been confronted with two baselines, *Clairvoyant*, offering always the top k most frequent results and *Zangarle*, a model proposed in a related work [50]. Our method is the most efficient for precision at one P@1, as can be seen in Table 7.4. Our competitor behaves better for prediction at

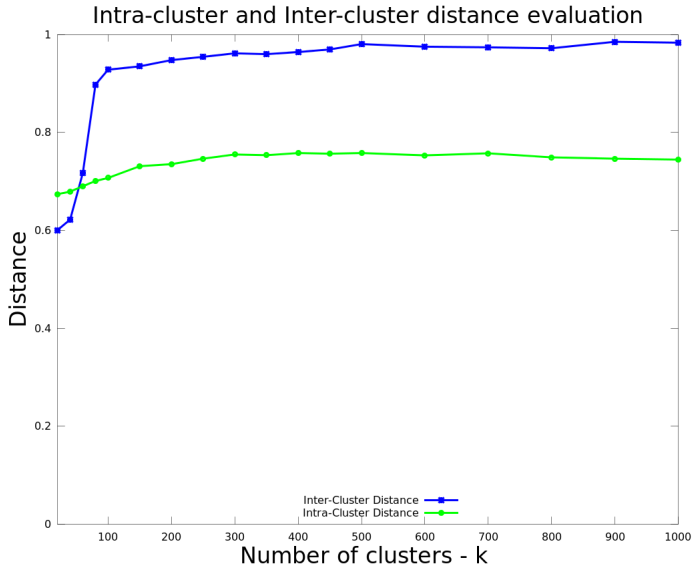

Figure 7.2: Evaluation of K-means for Dataset14 by varying k

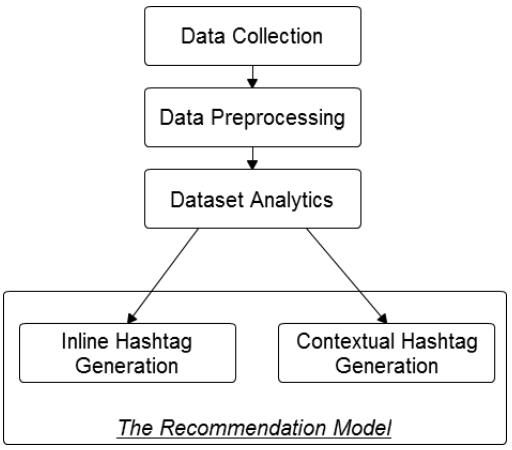

Figure 7.3: #SHARP! architecture

	Measure	*Clairv.*	*Zangerle*	#SHARP!				
				$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
	Precision	0.0001	0.116	0.107	**0.125**	**0.121**	0.092	0.085
	Recall	0.0001	0.092	0.085	**0.099**	**0.094**	0.069	0.064
@1	F1	0.0001	0.099	0.091	**0.106**	**0.101**	0.075	0.069
	F2	0.0001	0.094	0.087	**0.101**	**0.096**	0.071	0.065
	NDCG	0.0001	0.116	0.107	**0.125**	**0.121**	0.092	0.085
	Precision	0.002	0.078	0.063	0.073	0.075	0.069	0.057
	Recall	0.004	0.176	0.143	0.162	0.167	0.153	0.127
@3	F1	0.003	0.103	0.083	0.095	0.098	0.090	0.075
	F2	0.003	0.135	0.109	0.124	0.128	0.118	0.097
	NDCG	0.001	0.092	0.073	0.084	0.085	0.075	0.063
	Precision	0.002	0.062	0.047	0.052	0.054	0.053	0.046
	Recall	0.009	0.228	0.173	0.191	0.198	0.195	0.168
@5	F1	0.003	0.093	0.070	0.078	0.081	0.079	0.069
	F2	0.005	0.141	0.106	0.118	0.122	0.120	0.103
	NDCG	0.002	0.081	0.059	0.067	0.068	0.062	0.054

Table 7.4: Performance of #SHARP! against the baseline (*Clairvoyant*) and a state-of-the-art competitor (*Zangerle*).

3 and at 5 (P@3 and P@5). We have used various evaluation measures for ranked and non-ranked results as well.

We also make an analysis of what can be improved in our methodology. We discover that inline recommendation performs better for big values of k, namely when offering more recommendations, while contextual for smaller values of k, when offering fewer recommendations. Due to the fact that Zangarle and contextual recommendations are conceptually similar, we test the performance of #SHARP! by replacing contextual recommendation with Similarity ranking by Zangarle. We see, in Figures 7.4 and 7.5 that the mix between two methods outperforms each individual methods, offering the better precision and recall than all the other cases.

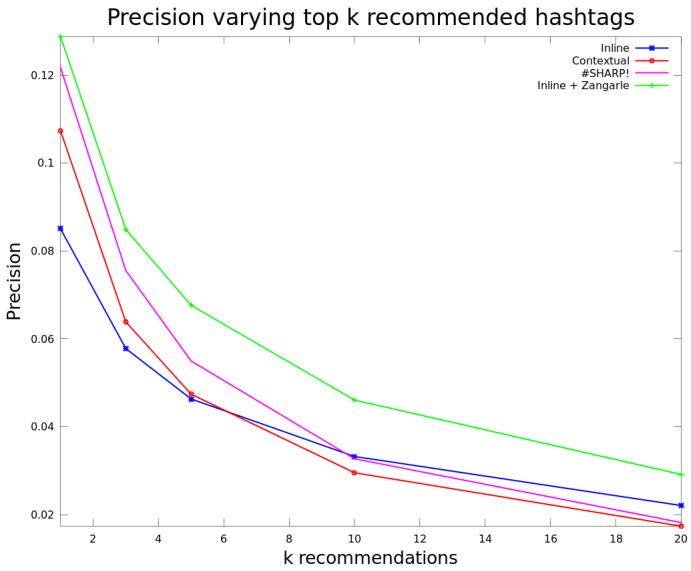

Figure 7.4: Precision @k

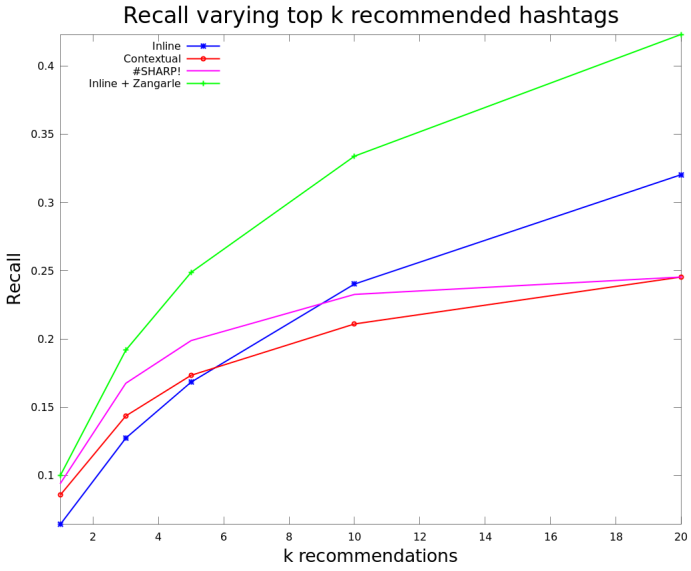

Figure 7.5: Recall @k

Chapter 8

Conclusions

8.1 Conclusions & future work

Keeping in mind the objectives we have previously set, we focused our attention on possible ways to approach those objectives and offer possible solutions. We have achieved the following:

- We offered an expanded definition of Social Media and aggregated opinions from other Social Media research. Due to the fact that Social Media has as a medium many social networking websites, we took a look at what social networks mean, when was the concept first used and how it has changed during time. We used Twitter as a case study and analyzed it from a double perspective: that of the user and that emerging from social network analysis.

- In order to understand the mechanisms behind Social Media platforms we investigated aspects regarding user experience and user engagement. An insight into the consumer psychology can offer more clues on how to stimulate engagement. Our proposed approach was exploiting the ludic tendency of consumers, namely how can products and services be transformed in order to appeal to the sense of playfulness of a user. Gamification is a techniques that tries to apply game mechanics to non-game contexts in order to stimulate and engage users. We made an overview of the basic principles behind it and see how they can actually be applied to an example application and how it is applied to a social media platform, Twitter. Results of this study were disseminated in [30].

- Advancing in our analysis of Twitter we look at the most important applications. Twitter is an interesting study as it shows different par-

ticularities. It is considered both a social network and a microblogging environment. If papers in the field of social network analysis deal with Twitter as a social network, papers in the field of text and short text analysis approach Twitter as microblogging platform. We explored the related work regarding Twitter and its characteristics and applications. We wanted to understand its structure and users by taking interest in social network analysis articles that study Twitter, understand the motivations of users using this platform, why and how they use it, understand its content and how information circulates in the network by taking interest in content analysis and information diffusion, and look at possible applications like predicting popular messages, short text classification, discovering news in tweets and so on.

- In order to analyze Twitter content we first needed to collect it. We explored the possible options: the classical option, by crawling the website and extract the content from the HTML pages, the methods were disseminated in [29, 27, 28] or the streaming option Twitter offers at our disposal, taking advantage of the powered stream for scientific research, ten times larger than the regular one, which provides sampled content in a JSON format. We chose to create a private connection to the Twitter stream and download data directly from the source, after requesting special access rights.

- After looking at the downloaded data, we explored methods for how to better process this big amount of data. We took a look at the possible technologies to use and decided to process data in a distributed way, by using frameworks build for parallel computing. We decide upon Cascading, a MapReduce abstraction library build over Hadoop. We implemented several preprocessing tasks in order to clean data, filter relevant data for our analysis and build statistics over the obtained dataset. A description of how these tools were used in order to process our data can be found in [31].

- Considering the data at our disposal we explored ways on how we can model the data according to our objectives. We studied relevant methodologies and algorithms from Information Retrieval, Recommendations and Machine Learning in order to further implement them accordingly for the proposed application. We decide on using: 1) both supervised and unsupervised learning algorithms - the machine learning approach, 2) create an inverted index in order to retrieve the desired data - the information retrieval approach and 3) create a predictor -

the probabilistic approach. Using these techniques we model our data in different ways, while the meta-problem remains the same.

- The applicative focus of our research is proposing a valid representation for Twitter hashtags. Using this representation we cluster hashtags in a various number of clusters and analyze how they behave according to cluster size. We observe the fact that hashtags have their particularities and often seem a puzzle. Our motivation was to find automatic ways of grouping them together according to context and grasp their semantic sense from the correlation with other hashtags and top frequent terms in clusters. This work has been disseminated in [31].

- We also proposed an application that wishes to recommend hashtags to users according to a tweet they input. We first proved the usefulness of hashtags and explained how they function in rising the visibility of a tweet. Then we observed the dual nature of hashtags and divide them into two types: inline and contextual. We modeled each type of hashtag differently. The contextual hashtags refered to the semantic sense of a tweet, so they grasp a category or a certain larger topic. We simulated a search engine for hashtags and index them as documents. We queried the index with various tweets and according to the similarity of that tweet with the virtual document of the hashtag and ranked the appropriate hashtags. The inline hashtags were modeled through a probabilistic approach. We built all the combinations of words from a tweet and look for the most probable hashtags according to the input, which we rank according to the probability score. We evaluated the results and confronted them with the state-of-the-art.

Future work

An interesting application in the case of this experiment, which we wish to try in future work to perform hierarchical clustering with the purpose of building a hashtag taxonomy, starting from general to specific. Another interesting application would be creating a tool that automatically generates a human readable explanation for the meaning of a hashtags, probably by using topic modeling and summarization algorithms.

We believe these results can be further improved. We believe recommendations can become more effective for the retrieval problem. We wish to experiment with other ways of representing hashtags. We also want to apply a learning algorithm for determining the optimum weights in the linear combination of inline and contextual recommendations. We have come up with various features that can be used to reinforce and deflate certain hashtags like the entropy of a hashtag or the probability of a contextual hashtag

given an inline one. We intend to use these features and apply learning to rank techniques on this set of results. We also want to experiment another recommendation model through classification. We intend to experiment classification by using Naive Bayes, Complementary Naive Bayes and Support Vector Machines (the state-of-the-art in text classification).

We conclude the research by saying that we have approached the proposed objectives and proposed solutions to them. We consider to have significantly improved our knowledge about Social Media and and how to cluster short text and offer recommendation, while also contributing with novel solutions: applications, models and methodologies. However, there is a lot to be discovered in this field of research.

8.2 Dissemination of results

The results presented in this work were disseminated through the following articles:

Published/accepted in proceedings of international foreign conferences:

- C. I. Muntean, G. A. Morar, and D. Moldovan, *Exploring the meaning behind Twitter hashtags through clustering*, 15th International Conference on Business Information Systems, Vilnius, Lithuania, In Business Information Systems Workshops, LNBIP, vol. 127, pages 231 - 242. Springer-Verlag Berlin, 2012. (accepted for publishing)

- C. I. Muntean. *Raising engagement in e-learning through gamification.* In Proceedings of the 6th ICVL Conference, pages 323-329. Editura Universitatii Bucuresti, 2011.

- C. I. Muntean, D. Moldovan, and O. Veres. *A data mining method for accurate employment search on the web.* In Proceedings of the 2010 international conference on COMATIA, pages 123-128, World Scientific and Engineering Academy and Society, 2010.

Published/accepted in journals with national scope:

- G.A. Morar, C.I. Muntean, and G.C. Silaghi. *Implementing and running a workflow application on cloud resources.* Informatica Economica, 15(3):15-27, 2011.

Published/accepted in international foreign journals:

- C. I. Muntean, D. Moldovan, and O. Veres. *A personalized classification of employment offers using data mining methods.* In International Journal of Mathematical Models and Methods in Applied Sciences, 5(4):525-532, 2011.

8.3 List of Articles

The complete list of papers developed during the doctoral studies were disseminated through the following articles:

Published/accepted in proceedings of international foreign conferences:

- C. I. Muntean, G. A. Morar, and D. Moldovan, *Exploring the meaning behind Twitter hashtags through clustering*, 15th International Conference on Business Information Systems, Vilnius, Lithuania, In Business Information Systems Workshops, LNBIP, vol. 127, pages 231 - 242. Springer-Verlag Berlin, 2012. (accepted for publishing)
- C. I. Muntean. *Raising engagement in e-learning through gamification.* In Proceedings of the 6th ICVL Conference, pages 323-329. Editura Universitatii Bucuresti, 2011.
- R. Baraglia, C. Frattari, C. I. Muntean, F. M. Nardini, and F. Silvestri. *RecTour: A recommender system for tourists.* In Proceedings of the 2012 Web Intelligence Workshops, WIIAT'12, Macau, China, 2012. IEEE Computer Society. (accepted for publishing)
- R. Baraglia, C. Frattari, C. I. Muntean, F. M. Nardini, and F. Silvestri. *A trajectory-based recommender system for tourism.* In Proceedings of 2012 International Conference on Active Media Technology, AMT'12 LNCS, Macau, China, 2012. Springer. (accepted for publishing)
- C. I. Muntean, D. Moldovan, and O. Veres. *A data mining method for accurate employment search on the web.* In Proceedings of the 2010 international conference on COMATIA, pages 123-128, World Scientific and Engineering Academy and Society, 2010.

Published/accepted in journals with national scope:

- G.A. Morar, C.I. Muntean, and G.C. Silaghi. *Implementing and running a workflow application on cloud resources.* Informatica Economica, 15(3):15-27, 2011.

- G. Morar, C. I. Muntean, N. Tomai, *An Adaptative M-learning Architecture for Building and Delivering Content based on Learning Objects*, The Second Romanian Workshop on Mobile Business, 10-11 Septembrie 2010. In Informatica Economica, Vol.10, No 1/2010, pp. 63-73.

Published/accepted in international foreign journals:

- C. I. Muntean, D. Moldovan, and O. Veres. *A personalized classification of employment offers using data mining methods.* In International Journal of Mathematical Models and Methods in Applied Sciences, 5(4):525-532, 2011.

Submitted for review:

- Claudio Lucchese, Cristina Ioana Muntean, Raffaele Perego, Fabrizio Silvestri, Hossein Vahabi, Rossano Venturini, *Recommendation Systems in UCG*, book chapter in Mining of User Generated Content and Its Applications to be published by Tailor & Francis (CRC Press).

Working paper:

- Diego Ceccarelli, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Fabrizio Silvestri, *#SHARP! : a System for HAshtag Recommendation.*

Bibliography

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[2] Albert Angel, Nick Koudas, Nikos Sarkas, and Divesh Srivastava. What's on the grapevine? In Ugur Çetintemel, Stanley B. Zdonik, Donald Kossmann, and Nesime Tatbul, editors, *SIGMOD Conference*, pages 1047–1050. ACM, 2009.

[3] S. Attfield, G. Kazai, M. Lalmas, and B. Piwowarski. Towards a science of user engagement (position paper). In *WSDM Workshop on User Modeling for Web Applications*, February 2011.

[4] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[5] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using wikipedia. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788, New York, NY, USA, 2007. ACM.

[6] Ranieri Baraglia, Claudio Frattari, Cristina Muntean, Franco Maria Nardini, and Fabrizio Silvestri. Rectour: A recommender system for tourists. In *Proceedings of the 2012 Web Intelligence Workshops*, WI-IAT'12, Macau, China, 2012. IEEE Computer Society. Accepted for publishing.

[7] Ranieri Baraglia, Claudio Frattari, Cristina Muntean, Franco Maria Nardini, and Fabrizio Silvestri. A trajectory-based recommender system for tourism. In *Proceedings of 2012 International Conference on Active Media Technology*, AMT'12 LNCS, Macau, China, 2012. Springer. Accepted for publishing.

[8] Richard Bartle. *Designing Virtual Worlds*. New Riders Games, 2003.

[9] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on Twitter. In *Proceedings of the Seventh Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, July 2010.

[10] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Structural trend analysis for online social networks. *Proc. VLDB Endow.*, 4:646–656, July 2011.

[11] P.J. Carrington, J. Scott, and S. Wasserman. *Models and methods in social network analysis*. Cambridge Univ Pr, 2005.

[12] D. Ceccarelli, Muntean C.I., Nardini F.M., Perego R., and Silvestri F. #sharp! : a system for hashtag recommendation. 2012. working paper.

[13] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1185–1194, New York, NY, USA, 2010. ACM.

[14] Cheng-Tao Chu, Sang Kyun Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Andrew Y. Ng, and Kunle Olukotun. Map-reduce for machine learning on multicore. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 281–288. MIT Press, Cambridge, MA, 2007.

[15] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[16] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.

[17] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. Gamification. using game-design elements in non-gaming contexts. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, pages 2425–2428, New York, NY, USA, 2011. ACM.

[18] Edd Dumbill. What is big data? `http://radar.oreilly.com/2012/01/what-is-big-data.html`, 2012.

[19] David R. Flatla, Carl Gutwin, Lennart E. Nacke, Scott Bateman, and Regan L. Mandryk. Calibration games: making calibration tasks enjoyable by adding motivating game elements. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 403–412, New York, NY, USA, 2011. ACM.

[20] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178, New York, NY, USA, 2010. ACM.

[21] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60:2169–2188, November 2009.

[22] J.H. Kietzmann, K. Hermkens, I.P. McCarthy, and B.S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 2011.

[23] D. Knoke and J.H. Kuklinski. *Network analysis*, volume 28. Sage Publications, Inc, 1982.

[24] C. Li and J. Bernoff. *Groundswell: Winning in a world transformed by social technologies*. Harvard Business School Press, 2011.

[25] C. Lucchese, Muntean C.I., Perego R., Silvestri F., Vahabi H., and Venturini R. Recommendation systems in ucg. *Mining of User Generated Content and Its Applications*, 2012. book chapter submitted for review.

[26] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[27] G.A. Morar, C.I. Muntean, and G.C. Silaghi. Implementing and running a workflow application on cloud resources. *Informatica Economica*, 15(3):15–27, 2011.

[28] C.I. Muntean, D. Moldovan, and O. Veres. A data mining method for accurate employment search on the web. In *Proceedings of the 2010 international conference on Communication and management in technological innovation and academic globalization*, pages 123–128. World Scientific and Engineering Academy and Society (WSEAS), 2010.

[29] C.I. Muntean, D. Moldovan, and O. Veres. A personalized classification of employment offers using data mining methods. *Internationa Journal of Mathematical Models and Methods in Applied Sciences*, 5(4):525–532, 2011.

[30] Cristina Ioana Muntean. Raising engagement in e-learning through gamification. Number 6 in 6th ICVL 2011, pages 323 – 329. Editura Universitatii Bucuresti, 2011.

[31] Cristina Ioana Muntean, Gabriela Andreea Morar, and Darie Moldovan. Exploring the meaning behind twitter hashtags through clustering. Number 127 in Lecture Notes in Business Information Systems, pages 231 – 242. Springer-Verlag Berlin, 2012.

[32] Brendan O'Connor, Michel Krieger, and David Ahn. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In William W. Cohen, Samuel Gosling, William W. Cohen, and Samuel Gosling, editors, *ICWSM*. The AAAI Press, 2010.

[33] A. Osterwalder. The business model ontology: A proposition in a design science approach. *Academic Dissertation, Universite de Lausanne, Ecole des Hautes Etudes Commerciales*, 2, 2004.

[34] A. Osterwalder, Y. Pigneur, et al. An e-business model ontology for modeling e-business. In *15th Bled Electronic Commerce Conference*, pages 17–19. Bled, Slovenia, 2002.

[35] A. Osterwalder, Y. Pigneur, and C.L. Tucci. Clarifying business models: Origins, present, and future of the concept. *Communications of the association for Information Systems*, 16(1):1–25, 2005.

[36] John Pavlus. The game of life. *Scientific American*, 303:43–44, 2011.

[37] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. 2011.

[38] E.T. Peterson and J. Carrabis. Measuring the immeasurable: Visitor engagement. *Research and Analysis from Web Analytics Demystified, the Web Analytics Thought Leaders*, 2008.

[39] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 91–100, New York, NY, USA, 2008. ACM.

[40] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 385–388, New York, NY, USA, 2009. ACM.

[41] C.K. Prahalad and V. Ramaswamy. *The Future of Competition: Co-Creating Unique Value With Customers*. Harvard Business School Pub., 2004.

[42] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51, New York, NY, USA, 2009. ACM.

[43] Don Tapscott and Anthony D Williams. *Wikinomics: How Mass Collaboration Changes Everything*, volume 58. Portfolio, 2006.

[44] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, IMC '11, pages 243–258, New York, NY, USA, 2011. ACM.

[45] M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding Connections on the Social Web*. Real Time Bks. O'Reilly Media, 2011.

[46] Alex H. Wang. Dont't Follow me: Spam Detection in Twitter. In *Proceedings of the International Conference on Security and Cryptography (SECRYPT)*, July 2010.

[47] 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.

[48] Tan Xu and Douglas W. Oard. Wikipedia-based topic clustering for microblogs. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011.

[49] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In William W. Cohen and Samuel Gosling, editors, *ICWSM*. The AAAI Press, 2010.

[50] E. Zangerle, W. Gassler, and Specht G. Recommending #-Tags in Twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web 2011*, pages 62–73. CEUR-WS, 2011.