Semantic Social Modeling of Online User-generated Content

Gerasimos Razis

Department of Computer Science and Biomedical Informatics

University of Thessaly

A thesis submitted for the degree of

Doctor of Philosophy

January 2019

Dedication

This thesis is dedicated to

my wife and daughter

Acknowledgements

First and foremost, I would like to thank my supervisor, Associate Professor Ioannis Anagnostopoulos, for the guidance, advice and trust that has shown me throughout this time, both on a professional and on a personal level. Moreover, I would like to thank the members of the committee of this thesis for their valuable advice. Finally, I would like to thank my parents and my family for their support during all stages of this effort.

Abstract

In this thesis, we study information management issues that arise in Online Social Networks (OSNs), as well as collective intelligence issues towards automated knowledge representation. We focus on three research directions, namely i) online social influence and the discovery of its impactful entities, ii) user-generated content and the role of semantics in social analysis, and iii) qualitative assessment of viral disseminated content. We present efficient and scalable methods focused on specific problems in the addressed directions, while we aim at proposing advancements in the relevant state of the art research in the field.

In particular, in the first research direction, we study how we can measure social influence and what are its application domains. To this end, we developed a service aiming at calculating and ranking the importance and influence of Twitter accounts. This service incorporates theoretical aspects of influence metrics that derive from social functions that evaluate i) the activity of a Twitter account (e.g. tweets, re-tweets and replies), ii) its social degree (e.g. followers and following) and iii) its network impact (e.g. content diffusion and social acknowledgement).

In the second research direction, we investigate the role of semantics in OSNs and the adoption of Semantic Web technologies which can be used for the detection of similar users, as well as user personalization issues (e.g. interests and suggestions). To this end, we define an ontological schema towards the semantification of social analytics, including structural aspects of Twitter accounts, disseminated entities and social relationships. Furthermore, we propose a methodology towards the discovery and suggestion of similar Twitter accounts, based entirely on their disseminated content. On top of that and based on the similarity relationships, we present an approach towards the automatic labeling of Twitter accounts by exploiting information from the Linked Open Data cloud; specifically according to DBpedia thematic categories. Finally, we contribute in the field of Query Expansion (QE) by proposing an algorithmic approach, which expands a user's query through the creation of a suggestion set that consists of the most viral and up-to-date Twitter entities.

Finally, in the third research direction, we tackle the problem of qualitative assessment of user-generated content by utilizing social influence and semantics. We conclude that the first two research areas along with the later can jointly provide useful insights, when we want to model dynamic properties of influential content and its flow dynamics.

Περίληψη

Η παρούσα διδακτορική διατριβή πραγματεύεται θέματα και προβλήματα διαχείρισης πληροφοριών που προκύπτουν εντός των Διαδικτυακών Κοινωνικών Δικτύων (Online Social Networks), καθώς και θέματα συλλογικής ευφυΐας προς την κατεύθυνση της αυτοματοποιημένης αναπαράστασης γνώσης. Σε αυτό το πλαίσιο ακολουθούνται τρεις ερευνητικές κατευθύνσεις συγκεκριμένα: i) η έννοια της επιρροής στα κοινωνικά δίκτυα και η ανεύρεση οντοτήτων σε αυτά με μεγάλη επιρροή, ii) το περιεχόμενο που δημιουργείται από τους χρήστες και ο ρόλος της σημασιολογίας στην ανάλυση των κοινωνικών δικτύων, και iii) η ποιοτική αξιολόγηση του διαχεόμενου περιεχομένου. Παρουσιάζουμε αποτελεσματικές και κλιμακώσιμες μεθόδους, εστιασμένες σε συγκεκριμένα προβλήματα των προαναφερθεισών κατευθύνσεων, με τελικό σκοπό να προταθούν νέες μέθοδοι στην αιχμή της έρευνας.

Στην πρώτη ερευνητική κατεύθυνση, μελετάμε πώς μπορούμε να μετρήσουμε την κοινωνική επιρροή και ποια είναι τα πεδία εφαρμογής της. Για το σκοπό αυτό, δημιουργήσαμε μια δημοσίως διαθέσιμη υπηρεσία με στόχο τον υπολογισμό και την κατάταξη της επιρροής και της επίδρασης λογαριασμών στο Twitter. Αυτή η υπηρεσία ενσωματώνει θεωρητικές πτυχές της μέτρησης επιρροής οι οποίες απορρέουν από κοινωνικές λειτουργίες που αξιολογούν i) την κοινωνική δραστηριότητα ενός λογαριασμού στο Twitter (π.χ. tweets, re-tweets, απαντήσεις), ii) την κοινωνική δημοτικότητα (π.χ. ακολούθους (followers), ακολουθούμενους (following)), και iii) τον αντίκτυπο στο κοινωνικό δίκτυο (π.χ. διάχυση περιεχομένου, κοινωνική αναγνώριση).

Στη δεύτερη ερευνητική κατεύθυνση, διερευνάται ο ρόλος της σημασιολογίας στα Διαδικτυακά Κοινωνικά Δίκτυα και η υιοθέτηση τεχνολογιών Σημασιολογικού Ιστού οι οποίες μπορούν να χρησιμοποιηθούν για την ανίχνευση παρόμοιων χρηστών καθώς και θέματα εξατομίκευσης χρήστη (π.χ. ενδιαφέροντα και προτάσεις). Υπό αυτό το πρίσμα, ορίζουμε ένα οντολογικό σχήμα με σκοπό τη σημασιολογική αναπαράσταση των αναλυτικών στοιχείων (analytics) των κοινωνικών δικτύων, συμπεριλαμβανομένων των δομικών πτυχών των λογαριασμών Twitter, των διαχεόμενων οντοτήτων, καθώς και των κοινωνικών σχέσεων. Επιπροσθέτως, προτείνουμε μια μεθοδολογία για την ανεύρεση και πρόταση παρεμφερών λογαριασμών στο Twitter, με βάση αποκλειστικά το διαχεόμενο περιεχόμενο. Συν τοις άλλοις και με βάση τις σχέσεις ομοιότητας, παρουσιάζουμε μια προσέγγιση για την αυτόματη σήμανση των λογαριασμών Twitter εκμεταλλευόμενοι πληροφορίες από το σύννεφο των «Συνδεδεμένων Ανοιχτών Δεδομένων» (Linked Open Data cloud), και συγκεκριμένα σύμφωνα με θεματικές κατηγορίες από τη γνωσιακή βάση DBpedia. Τέλος, συμβάλλουμε στο πεδίο της Επέκτασης Ερωτημάτων (Query Expansion) προτείνοντας μια αλγοριθμική προσέγγιση, η οποία επεκτείνει το ερώτημα ενός χρήστη μέσω της δημιουργίας ενός συνόλου προτάσεων το οποίο αποτελείται από τις πιο δημοφιλείς και ενημερωμένες οντότητες του Twitter.

Τέλος, στην τρίτη ερευνητική κατεύθυνση, αντιμετωπίζουμε το πρόβλημα της ποιοτικής αξιολόγησης του περιεχομένου που παράγουν οι χρήστες χρησιμοποιώντας την κοινωνική επιρροή και τη σημασιολογία. Καταλήγουμε στο συμπέρασμα ότι οι δύο πρώτες ερευνητικές περιοχές μπορούν από κοινού με την τρίτη να παράσχουν χρήσιμες πληροφορίες, όταν θέλουμε να αναπαραστήσουμε τις δυναμικές ιδιότητες του περιεχομένου με που έχει μεγάλο αντίκτυπο καθώς και της δυναμικής του ροής.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Nowadays, hundreds of millions of messages are shared on a daily basis among the users of Online Social Networks (OSNs). These users vary from citizens to political persons and from news agencies to large multinational corporations. In this "ocean" of information, a challenging task is the discovery of the important actors who are able to influence others and produce messages of high social quality, importance and recognition. Those influential users are also called *opinion leaders* (Riquelme and Cantergiani, 2016), *domain experts* (Liu et al., 2014), *influencers* (Razis and Anagnostopoulos, 2014a), *innovators* (Chai et al., 2013) and *prestigious* (Gayo-Avello, 2013) or *authoritative actors* (Bouguessa and Romdhane, 2015). Often, their degree of influence is also complemented or affected by various quality measurements which are based on their social semantics. The latter can either be related to the content of the messages (e.g. keywords, hashtags) or to the metadata of the user (e.g. activity, relationship details); very often Semantic Web technologies are employed for the transformation of unstructured data into Linked Data.

The exploitation of viral user generated content and the comprehension of the role and effect of influential nodes on flow dynamics have a huge potential to create insights and additional value across several domains, such as marketing, information retrieval, recommendation systems, community and/or event detection, query expansion, thematic categorization, homophily tendency, and sentiment analysis. As the OSNs' data volume and users' activity rapidly increase, complex challenges and problems start to emerge. In this thesis, we study the issues that stem from the overwhelming amount of information disseminated in OSNs, centering on the case of Twitter, and focusing on three major directions, namely on i) *social influence*, ii) *social semantics*, and iii) *qualitative assessment*. Our main aim is to provide efficient methodologies and techniques that advance the state-of-the-art in a representative set of problems that stems from each of these directions.

The first direction is concerned with the discovery of influential entities in OSNs, namely the important actors who are able to influence others and produce messages of high social quality, importance and recognition. In the related literature there is no solid agreement on what is meant by an influential user. Therefore, the term "influence" has multiple interpretations and every time it is considered in a different way. Consequently, a variety of influence measures is constantly emerging, while each of them is based on different criteria. Thus, efficient, well-defined, adaptable and extensible methods are needed in order to tackle these aspects.

The second direction is concerned with the role of semantics in modeling OSNs information and the challenges that arise when exploiting them towards a numerous set of application domains, such as for detection of similar users and communities, user personalization (e.g. interests, suggestions, etc.), topic identification, recommendation systems, and transformation of unstructured data into Linked Data. Specifically, handling raw unstructured or semi-structured data combined with the lack of a unified representation (e.g. ontological schemes) can lead to a series of issues, such as data sparseness, semantic gap, computational overhead or multiple interpretations of the same concepts. Moreover, social semantics representation using Semantic Web technologies can facilitate the linkage with the Linked Open Data (LOD) cloud, thus leading to enriched information and increased data value, which in turn can be used in classification or recommendation tasks. Finally, the widely-

adopted by search engines service of query suggestion which is based on the submitted by users' search queries fails to timely propose newly emerged and viral content. Thus, according to the aforementioned, the need arises for novel methodologies and techniques for retrieving, transforming and exploiting social semantic data.

The third and last direction is concerned with the qualitative assessment of viral user-generated content in OSNs, as well as with techniques for modeling the dynamic properties of influential content and its flow dynamics. The quality of that content varies from excellent and interesting to abusive and spam. As the availability of social media content increases, the task of identifying high-quality content based on user contributions, actions and preferences becomes increasingly important. Thus, efficient methods are needed in order to tackle these issues.

1.1. Contributions

In this thesis, we study the aforementioned directions and present efficient and scalable methods applied in OSNs for measuring influence, analyzing the role of semantics, and assessing the quality of the disseminated content, with the aim to provide targeted research advancements in the state of the art of the series of problems that are discussed herein. Specifically, in the first direction, we propose a novel influence metric deriving from a social function, aiming at calculating and ranking the importance and influence of Twitter accounts. In the second direction, we propose a novel ontological schema towards the semantification of social analytics and of structural relationships, as well as a methodology towards the discovery of similar Twitter accounts. Based on the latter, we present an approach towards the automatic labeling of Twitter accounts following the DBpedia thematic categories. Furthermore, we contribute in the field of query expansion by proposing an algorithmic approach which expands a user's query by creating a suggestion set consisting of the most viral and up-to-date Twitter entities. Finally, in the third direction, we tackle the problem of qualitative assessment of user-generated content, by introducing a novel qualitative factor. Specifically, our contributions include the following:

1) The growing volume of the disseminated content in OSNs has brought forth significant challenges regarding their exploitation, in a wide range of application domains. Therefore, we conduct a review covering two major aspects of OSNs, namely the online social influence and the role of semantics, while discussing how we can combine both aspects towards the qualitative assessment and modeling of user-generated content. We present in detail the methodologies as described in the most up-to-date and impactful studies relevant to the aforementioned aspects. This work addresses the problems of all these three directions tackled in this thesis. The discussed methods and the drawn conclusions appear in (Razis et al., 2018).

2) On a daily basis, hundreds of millions of messages are disseminated in Twitter by numerous accounts held by citizens, political persons, news agencies and large multinational corporations. Consequently, the discovery of the important actors who are able to influence others and produce messages of high social quality, importance and recognition becomes a challenging task. To this end, we present *"InfluenceTracker"*, a service aiming at calculating and ranking the importance and influence of Twitter accounts. Specifically, we propose a

novel measurement, namely the "*Influence Metric*", value of which derives from a social function incorporating i) the activity of a Twitter account (e.g. tweets, retweets, and replies), ii) its social degree (e.g. followers, and following) and iii) its qualitative content, reflected by other users' acknowledgement (e.g. retweets) and preferences (e.g. favorites). We perform an experimental evaluation on real world data and validate that the aforementioned social properties and characteristics are the appropriate ones for such a measurement to be based on. Moreover, by comparing our metric against a commercial one, we observe that ours is much more accurate and well-defined. This work addresses the problems of the first and the third of the three directions tackled in this thesis. The discussed methodologies, proposed algorithms and obtained results have been published in (Razis and Anagnostopoulos, 2014a), (Razis and Anagnostopoulos, 2014b), and (Razis and Anagnostopoulos, 2016).

3) The content found in OSNs can be characterized as highly unstructured, and also suffering from typographical errors, informal language, and high contextualization. Consequently, microblogging retrieval systems suffer from the problems of data sparseness and of semantic gap. In this context, we present an ontological schema towards the semantification of social analytics. Specifically, the proposed *"InfluenceTracker Ontology"* is capable of modeling structural aspects of Twitter accounts, including information of their owners, all of their disseminated entities (mentions, replies, hashtags, photos, and URLs), as well as their online social relationships, interactions (mentions, replies), and qualitative measurements. The structured semantified data are publicized through a SPARQL endpoint. In order to provide a five-star data model, according to Tim Berners-Lee's Linked Open Data (LOD) rating system (Berners-Lee, 2012), we extended our ontological schema by incorporating properties from DBpedia, and other ontologies. Since the latest update of the LOD cloud, on 20/02/2017, the *"InfluenceTracker"* dataset is officially part of this interlinked and interdependent ecosystem of data. Our system demonstrates the benefits of the increased value of the available data, by providing answers to sophisticated queries (e.g. return the top-10 members of political parties according to their "Influence Metric" value). To the best of our knowledge, there is currently no active service for providing such kind of data linkage, i.e. social analytics with the LOD cloud. By further exploiting the benefits of the semantic technologies and the LOD cloud, we propose the *"Thematic Category Labeling Algorithm"* to achieve an automatic labeling of Twitter accounts with respect to thematic categories derived from the properties of DBpedia knowledge base. Based on that semantified content, we introduce a methodology for discovering and suggesting similar Twitter accounts, in terms of interests, based entirely on their disseminated entities. The existence of an ontological scheme and the adoption of semantics technologies reduce the complexity of storing and retrieving specific segments of data and decrease the number of the necessary calculations required for the computation of the coefficients and metrics. This work addresses the problems of the second of the three directions tackled in this thesis. The discussed methodologies and proposed algorithms have been published in (Razis and Anagnostopoulos, 2014b), (Razis et al., 2015), (Razis et al., 2016), and (Razis and Anagnostopoulos, 2016).

4) Despite the fact that the set of social semantics of each account in an OSN is unique, as these semantics depend on personal social activities, common properties and patterns can be recognized among them. These can be exploited towards the discovery of users' social behavioral patterns, interests, and preferences. In this context, and under the assumption that the more common social entities are found in the disseminated messages of OSN accounts, the more similar, in terms of content or interest, they tend to be, we present the *"Similarity Metric"*, a novel measurement for identifying and suggesting similar Twitter accounts. Moreover, we describe an iterative algorithm towards the automatic labeling of Twitter accounts with respect to DBpedia thematic categories. In order to overcome the problem of automatically relating hundreds of Twitter accounts to DBpedia resources, we propose two generic and adaptable methodologies, which are evaluated against real-world data. This work addresses the problems of the second of the three directions tackled in this thesis. The discussed methodologies, proposed algorithms and obtained results have been published in (Razis and Anagnostopoulos, 2016), and (Razis et al., 2016).

5) There are many web information management methods and techniques that help search engines and news services to provide useful suggestions with respect to queries, thus facilitating the users' search. However, the penetration of microblogging services in our daily life demands to also consider the social sphere as far as query suggestions are concerned. Towards this direction, we introduce an algorithmic approach capable of creating a dynamic query suggestion set, which consists of the most viral and trendy Twitter entities (e.g. hashtags, user mentions, URLs) with respect to a user's query. Such content can be rapidly disseminated, as it is maintained and reproduced many times by multiple Twitter accounts along with other entities, thus generating a dynamic network of resilient content. We perform an experimental evaluation on real use-cases and we provide comparative results showing that our proposal outperforms other popular services or methods and baselines presented in the literature. This work addresses some of the problems of the second of the three directions tackled in this thesis. The discussed methodologies, proposed algorithms and obtained results have been published in (Anagnostopoulos et al., 2013) and (Anagnostopoulos et al., 2015).

1.2. Thesis Outline

The remainder of this thesis is structured as follows. Chapter 2 presents a systematic review of online social influence metrics, properties, and applications as well as the role of semantics in modeling OSNs information. Chapter 3 discusses the properties and presents a novel framework for measuring social influence. Chapter 4 proposes an ontological schema towards the semantification provision of Twitter analytics as well as methodologies based on Semantic Web technologies for data enrichment, and user classification. Chapter 5 presents novel frameworks for discovering similar Twitter accounts and interlinking social entities to the LOD cloud. Chapter 6 presents a query suggestion method which is based on viral and up-to-date Twitter entities. Finally, Chapter 7 concludes the thesis.

Chapter 2. Influence, Social Networks and Semantics

2.1. Introduction

The discovery of influential entities in all kinds of networks (e.g. social, digital, or computer) has always been an important field of study. In recent years, Online Social Networks (OSNs) have been established as a basic means of communication where very often influencers and opinion makers promote politics, events, brands or products through viral content. In this chapter, we present a systematic review on i) online social influence metrics, properties, and applications and ii) the role of semantic in modeling OSNs information. We found that both areas can jointly provide useful insights towards the qualitative assessment of viral user-generated content, as well as for modeling the dynamic properties of influential content and its flow dynamics.

We study two major aspects of OSNs, namely the online social influence (Section 2.4) and the role of social semantics (Section 2.5) in OSNs, towards the qualitative assessment of viral user-generated content (Section 2.6). Specifically, we examine how influence can be measured or predicted and what kinds of methodologies are used to measure influence (e.g. based on topology, diffusion or social authority), and what are the application domains. Regarding the role of semantics in OSNs, we analyze related works based on Semantic Web technologies along with network theory and graph properties for topic identification, detection of similar users and communities, as well as for user personalization (e.g. interests, suggestions, and so on).

To summarize, this chapter provides the following contributions:

1. We present a literature review which aims at helping both researchers and data scientists to better understand how
    a. viral content is propagated,
    b. the role and effect of influential nodes in its diffusion, and
    c. how we can measure the influence of users in social networks.
2. We describe the reasons why the proper use of semantics for users and their generated content can provide useful insights and qualitative conclusions for numerous domains such as marketing, information retrieval, recommendation systems, community and/or event detection, query expansion, thematic categorization, homophily tendency and sentiment analysis.
3. We propose a hierarchical classification scheme in order to adequately cover all the perspectives of the aforementioned aspects.

The remainder of the chapter is organized as follows. The next section describes the methodology approach we followed for conducting our literature review. Section 2.3 presents the related literature, stressing out the differentiation and the added value of this review. Section 2.4 defines online social influence and its effects in user generated content. In Section 2.5 we analyze the role of semantics and why they are significant if one wants to receive valuable and tangible insights among users and

their social communities. Section 2.6 presents related qualitative assessments for modeling the dynamic properties of influential content. Finally, Section 2.7 concludes this chapter.

2.2. Methodological approach

In order to perform a more detailed analysis and to adequately cover all perspectives of the aforementioned two aspects, we analyzed the reviewed related research works according to the hierarchical classification scheme depicted in Figure 2.1. In most of the cases, a referred work does not fall with the scope of only one topic, thus demonstrating that related research efforts in these fields are complementary.

More specifically, with respect to the online social influence, we classify the related works according to the following four topics.

- Topic 1 - Influence Metrics: This topic includes works proposing methodologies that define online social influence and how to measure it. Thus, this topic is further divided into three subtopics namely a) Direct social information-based metrics, b) Hyperlink-based metrics, and c) metrics based on Machine Learning techniques.
- Topic 2 - Information Flow and Influence: This topic examines the impact of users with respect to viral properties of information as well as information propagation and information diffusion. Although there is no clear distinction between ‘propagation’ and ‘diffusion’ in the literature covering the OSNs and often these terms are used interchangeably, in this survey we explicitly examine separately the impact of influence in information propagation and on information diffusion. Diffusion relates to the spread of information from a starting node toward the rest of the network, while propagation takes into consideration the intermediate nodes as well, which receive, process, and further decide whether to re-transmit, re-direct or block the information. Thus, in this document we divide the information flow and influence topics into two subtopics namely, Propagation-oriented and Diffusion-oriented.
- Topic 3 - Network / Graph Properties: This category contains works which utilize the topology of a network or its structure in order to measure influence. Usually only a fraction of the whole network is used due to hardware or complexity limitations.
- Topic 4 - Applications: This topic presents the usage of the above metrics, mainly in applications that provide solutions for opinion makers, data analysts and information scientists. This topic is further divided into three subtopics namely, a) Ranking, b) Recommendation and c) Other application domains, such as sentiment analysis, and event detection, .

As for the role of social semantics in the provision of a qualitative assessment of viral user-generated content, we classify the related works we have reviewed into three topics.

- Topic 1 - Social Modeling: This topic contains approaches that adopt semantics for modeling the logical topology and structure of online social networks and media as well as the disseminated information.
- Topic 2 - Social Matching: The studies presented on this topic exploit the use of social semantics for identifying similar properties and activities with respect to user-generated content, description of real-life events, as well as revealing user interests and behavioral patterns across different online social media users. We divide this topic into two subtopics, namely a) User-oriented (e.g. similar user recommendation, user preferences, and so on), and b) Topic and Event-oriented (e.g. topic profiling and user interest, event detection, product marketing).
- Topic 3 - Community Detection: This category covers works that use social semantics for the detection of communities in OSNs.

To conduct our literature review, we collected 126 studies strongly related to the aforementioned issues. Initially, we used a specific set of related keywords (some indicative keywords are depicted in Table 2.1) as input for the discovery of relevant publications by submitting them through the academic digital library and search engine Application Programming Interfaces (APIs). Specifically, we utilized open access repositories (Google Scholar[1], arXiv[2], SSRN[3], and PLOS ONE[4]) and digital libraries that request subscription (ACM Digital Library[5], IEEE Xplore[6], and Elsevier[7]). Afterwards, we performed a review of the selected studies to highlight the most relevant topics and subtopics related to the influence in OSNs and social semantics. In the final step, we further filtered the selected works based on their date of publications, thus keeping the most recent. However, in order not to exclude the older but significant related works (with high citation counts), we described their impact in the newer works that have cited them. In this way, we kept our selected works up-to-date. The selected publications consist of three types, namely peer-reviewed journals, international conferences and workshops, and white papers in acknowledged academic repositories and archives. Figure 2.2 shows the distribution of the selected works in terms of their publication year and type. More than half of the publications have been published after 2014. The distribution according to their publication type is 37% journals, 60% conferences, and 3% white papers.

[1] https://scholar.google.com
[2] https://arxiv.org/
[3] https://www.ssrn.com/
[4] https://journals.plos.org/plosone/
[5] http://dl.acm.org/
[6] http://ieeexplore.ieee.org/Xplore/home.jsp
[7] https://www.elsevier.com/catalog?producttype=journals

Table 2.1: Indicative keywords used to search appropriate publications (many were used combined with the "AND" Boolean operator in conjunction with terms such as OSNs, online social networks, social media, and so on)

Influence	Social semantic modeling	Tweet quality
Influence maximization	Context-dependent influence	Event detection
Influence propagation	Content-driven approach	Information quality
Information propagation	Diffusion	Query expansion
Social network semantics	Sentiment-based influence	Social information retrieval
User interest	Similarity	Social recommendation

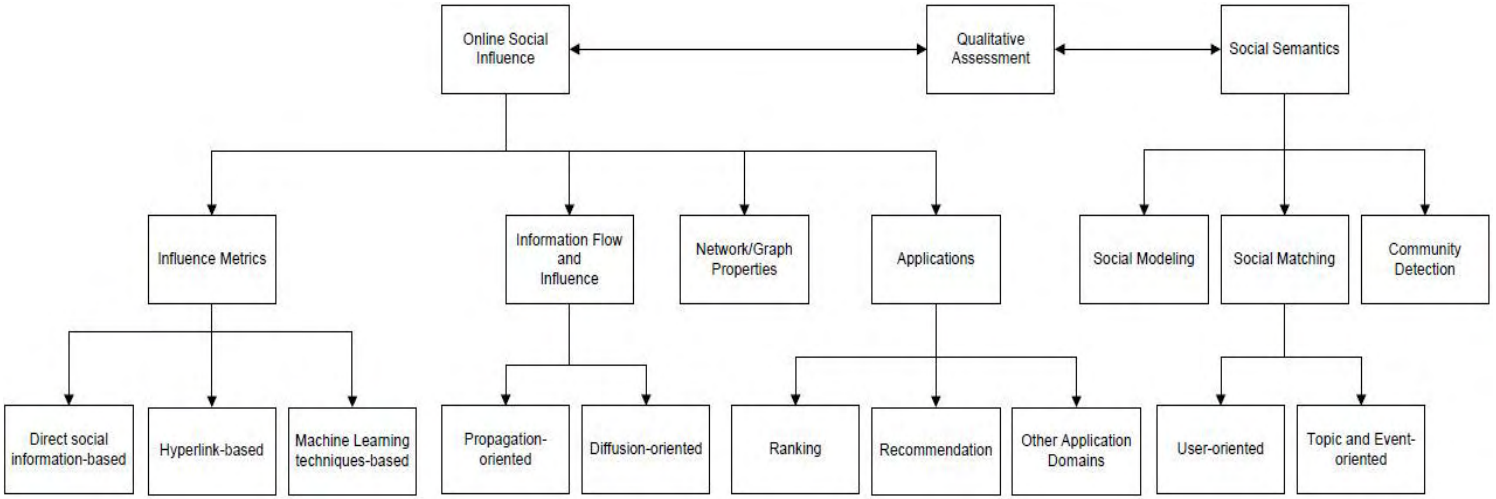

Figure 2.1: The hierarchical classification scheme followed in this work

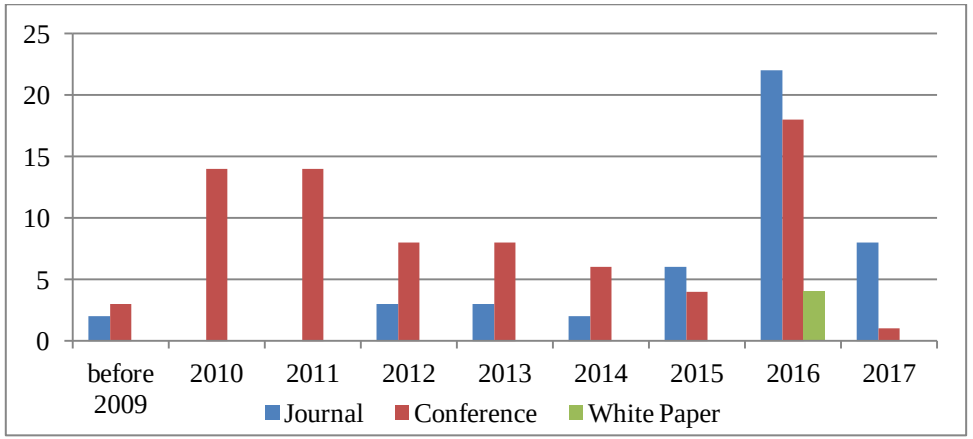

Figure 2.2: Distribution of publications per year for the selected works using IEEE Xplore, ACM Digital Library, Elsevier, Google Scholar, arXiv, SSRN, and PLOS ONE

2.3. Related Work

In this section, we present other survey papers from the related literature that tackle similar issues in terms of modeling online social influence with semantics. Then, we highlight the differentiation and the added value of this review, as well as our contributions across our classification scheme.

2.3.1. Similar Surveys

As already mentioned, the reviewed works were classified into a hierarchical scheme. This scheme is depicted in Figure 2.1, resulting in 20 hierarchically structured categories. For the purposes of this extensive review, we also considered other survey papers that tackle similar aspects in terms of the impact of influence in OSNs and the role of semantics ((Riquelme and Cantergiani, 2016), (Singer, 2016), (Bouadjenek et al., 2016), (Kumar et al., 2016) and (Bai et al., 2015)).

The authors in (Riquelme and Cantergiani, 2016) focused mainly on the classification of current diverse measurements aimed at discovering influential users in Twitter. Their range varies from those based on simple metrics provided by the Twitter API to the adoption of the PageRank algorithm and its variations. Other important factors are the content of the messages as some are focused on specific topics, their quality, in terms of likeability by others, as well as the activity and popularity of the users. Four of the aspects of our suggested scheme were covered, namely "Influence Metrics", "Network/Graph Properties", "Social Matching: Topic and Event-oriented", and "Qualitative Assessment".

In (Kumar et al., 2016), the authors analyzed a variety of OSN-based measurements and examined factors capable of affecting user influence. Thee metrics were grouped under various criteria deriving from:

- Neighborhood attributes, including the number of influencers, and exposure to direct and indirect influence.
- Structural diversity metrics that quantify the activity of the communities.
- Influence of locality and decay.
- Temporal measures including time delay until the reposting of a message.
- Cascade-based criteria, including its size and path length of messages.
- Metadata existence, including the presence of links, mentions, or hashtags.

Moreover, experiments were performed to predict user influence by using machine learning algorithms, with the aforementioned measurements as features. Based on our classification of this work, the survey described in (Kumar et al., 2016) covers the "Applications: Ranking" and "Network/Graph Properties" categories.

The work in (Singer, 2016) presents an overview of studies regarding Adaptive Seeding (AS) methodologies to solve the Influence Maximization (IM) problem. IM is the process of discovering and activating a set of seed influential nodes-users to initiate the diffusion process so that the largest number of nodes is reached or influenced. Often, this set of users is restricted to these ones who are engaged with the topic of interest, and due to structural dependencies of the network, it is possible to

rank low in terms of their influence potential. An alternative approach is to consider an adaptive method which aims at seeding neighboring nodes of high influence. As both the IM and the AS methodologies include the activation of nodes which in turn propagate the received information and activate others, the work described in (Singer, 2016) covers only the "Information Flow and Influence: Propagation-oriented" category as described in our survey.

The authors in (Bouadjenek et al., 2016) have reviewed approaches that enable the Information Retrieval (IR) tasks in OSNs, which exploit content and structural social information. The research works the authors have reviewed have been classified into three categories according to the use of social information. Specifically, the "social web search" category includes techniques where the social content is used to improve the classic IR processes such as the re-ranking of the retrieved documents, query reformulation, expansion or reduction, and user profiling. The second category, called "social search", includes methodologies on information discovery based on the users' generated content, interactions, and relationships. Finally, the "social recommendation" aims at predicting users' interests and is based on content-based and collaborative filtering approaches. Hence, the survey in (Bouadjenek et al., 2016) covers the aspects of "Social Matching: User-oriented", "Network/Graph Properties", and "Applications: Recommendation", as described by this work.

Finally, the authors of (Bai et al., 2015) provide an overview on various user classification methodologies in OSNs. More specifically, they describe the most common frameworks based on machine (i.e. Bayesian, Decision Tree, Logistics, SVM and KNN) and non-machine (concept of entropy and based on user similarity) learning techniques. The aim of these methodologies is to classify users into certain categories according to their explicit or implicit features, such as behavioral attributes, profile information, interests, viral content and interactivity. As a result, it covers only the "Social Matching: User-oriented" category as presented in this survey.

2.3.2. Review Differentiation and Extension

Table 2.2 provides a comparative evaluation of the surveys described in (Riquelme and Cantergiani, 2016), (Singer, 2016), (Bouadjenek et al., 2016), (Kumar et al., 2016) and (Bai et al., 2015). It consists of three columns. For every single, the first two represent the category and sub-category according to our classification scheme (Figure 2.1), as well as the respective section where we analyze it. The third column depicts the respective reference. The mark "✖" is placed in case where -according to the best of our knowledge- there is no other similar survey that covers this category.

Table 2.2: Classification of referenced surveys

Category / Subcategory		Section	Reference
Influence Metrics	*Direct social information-based*	2.4.1.1	(Riquelme and Cantergiani, 2016)
	Hyperlink-based	2.4.1.2	(Riquelme and Cantergiani, 2016)
	Machine Learning techniques-based	2.4.1.3	(Riquelme and Cantergiani, 2016)
Information Flow and Influence	*Propagation-oriented Approaches*	2.4.2.1	(Singer, 2016)
	Diffusion-oriented Approaches	2.4.2.2	×
Network / Graph Properties		2.4.3	(Riquelme and Cantergiani, 2016)
			(Bouadjenek et al., 2016)
			(Kumar et al., 2016)
Applications	*Ranking*	2.4.4.1	(Kumar et al., 2016)
	Recommendation	2.4.4.2	(Bouadjenek et al., 2016)
	Other Application Domains	2.4.4.3	×
Social Modeling		2.5.1	×
Social Matching	*User-oriented*	2.5.2.1	(Bouadjenek et al., 2016)
			(Bai et al., 2015)
	Topic and Event-oriented	2.5.2.2	(Riquelme and Cantergiani, 2016)
Community Detection		2.5.3	×
Qualitative Assessment		2.6	(Riquelme and Cantergiani, 2016)

Thus, compared to the surveys presented in Section 2.3.1, this review aims at covering and analyzing four additional aspects of OSNs, namely ”Information Flow and Influence” (further categorized in “Propagation-oriented” and “Diffusion-oriented”), “Social Modeling”, “Community Detection”, as well as ”Other Application Domains” related to online influence (Table 2.2). Moreover, the differentiation and extra issues covered in this work can be summarized in the following points:

- *Information Flow and Influence*: In contrast to many research works where the terms "diffusion" and "propagation" are used interchangeably, we tried to explicitly differentiate them by providing a clear distinction between their impact and their role in the disseminated information.
- *Social Modeling*: We include studies aiming at the transformation of unstructured social data into Linked Data, by i) relating entities to knowledge bases (e.g. Google Knowledge Graph, DBpedia), and ii) representing them as concepts extracted from ontologies using semantic vocabularies.
- *Community Detection*: We consider approaches that also employ social semantics and ontologies. Such approaches are not only useful for the analysis of OSNs, but also for understanding the structure and the properties of complex networks.
- *Other Application Domains*: We also consider additionally topics and approaches that exploit social influence for analyzing sentiment and user polarity, as well as the detection of critical real life events.

2.4. Online Social Influence

In this section, we describe one major aspect in OSNs, namely the online social influence. We focus on how influence can be measured or predicted and the methodologies (e.g. based on topology, diffusion or social authority) that can be used to measure influence along with the respective application domains.

2.4.1. Influence Metric

The calculation of the impact a user has on social networks, as well as the discovery of influencers on them is not a new topic. It covers a wide range of sciences, ranging from sociology to viral marketing and from oral interactions to OSNs. In the related literature there is no strong agreement on what is meant by the term "influential user". Therefore, the term "influence" has multiple interpretations. Consequently, emerging influence measures are constantly varying with each of them using different criteria. Despite this variation, all the related studies share a common result, which is that the most active users or those having the most followers are not necessarily the most influential ones. The works presented in this section discuss issues related to the discovery of influence and we classify these issues into three categories according to the way they a) exploit the direct social information (number of followers, followees, social content,), b) incorporate PageRank and related hyperlink-based algorithms, and c) employ machine learning techniques.

2.4.1.1. Direct Social Information Metrics

The study in (Anger and Kittl, 2011) proposes the "Social Networking Potential" as a quantitative measurement for discovering influential users in Twitter, and suggests that having a large number of followers does not guarantee a high influence. Their methodology is based on the number of tweets, replies, retweets, and mentions of an account.

The authors in (Cha et al., 2010) introduce three types of influence, namely "In-degree" (number of followers), "Retweet" (number of user generated tweets that have been retweeted) and "Mention" influence (number of times the user is mentioned in other users' tweets) for Twitter users. A necessary condition for the computation of these influence types is the existence of at least ten tweets per user. The authors claim that "Retweet" and "Mention" influence correlate well with each other, while the "In-degree" does not. Therefore, they conclude that the most followed users are not necessarily influential

Influence in terms of activity or passivity for Twitter users is studied in (Romero et al., 2011a). To conduct this study, a large number of tweets are utilized containing at least one URL, their creators and their followers. The influence metric produced depends on the "Follower-Following" relations of the users, as well as their retweeting behavior. As most studies in this area, it is stated that the number of followers a user has is a relatively weak predictor of the maximum number of views a URL can achieve.

In (King et al., 2013) the "t-index" metric is proposed, which aims at measuring the influence of a user on a specific topic. It is also based on the h-index factor and denotes the number of times a user's tweet on a specific topic has been retweeted. The authors suggest that a high influence on one topic does not necessarily mean the same on other topics.

A framework exploiting influence for evaluating and enhancing communication issues between governmental agencies and citizens in OSNs is proposed in (Dennett et al., 2016). The aim here is to evaluate the quality of the agencies' responses with respect to the citizens' requests, to analyze the citizens' sentimental attitude and their subsequent behaviors, and to suggest influential users to the agencies in order to obtain new audiences. To achieve these, several components are incorporated into the framework, which detect the demographics of the followers, their locations, topics of interest, and sentiments.

The authors in (Hassan et al., 2016) propose a different kind of influence called the "susceptibility to influence". Its metric estimates how easily a Twitter user can get influenced. The proposed metric utilizes the user's social interactions that depend on three factors namely, activity, sociability and retweeting habit. The activity reflects the user's tendency to interact with friends and, consequently, the chance to become influenced by them, while the "sociability" corresponds to the users' social degree among their activities, implying that interactions with more friends result in a wider diversity of topics and interests.

Finally, the study in (Peng et al., 2017) presents a methodology for measuring social influence in mobile networks by incorporating the entropy of the influence. Specifically, the friend and the interaction frequency entropies are introduced in order to describe the complexity and uncertainty of social influence. A weighted network is constructed based on the users' interactions, upon which three types of influence are introduced, namely a) direct influence among related users, b) indirect influence among unrelated users, and c) global influence that covers the whole network.

2.4.1.2. Hyperlink-based Metrics

The studies in this sub-section describe influence metrics through hyperlink-based algorithms (e.g. PageRank).Therefore, in this approach the influence is strongly related to the structure and the topology created by the OSN itself.

The authors in (Wei et al., 2016b) propose an influence model based on two aspects, user relationship and activity. They used three factors namely Influence Diffusion Model (IDM), PageRank, and usage behavior. IDM focuses on tweets and their reply chain, while providing the influence of propagation based on word occurrence. The PageRank algorithm is employed for calculating users' significance based on their relationships, while the user behavior factor affects a user's influence score as it is based on the number of posts, mentions, followers, and retweets. The core ideas of these models are extracted and are integrated into proposed influence model.

An influence ranking method is proposed in (Li et al., 2013b) based on the fact that the influence of a user is determined by the followers' influence contribution, which in turn highly depends on their interactions. A user can exert more influence over another if the former writes more tweets related to those of the second user. The proposed measurement is a variation of the PageRank algorithm and is based on a similarity factor between published the tweets over a graph of following users, retweets, mentions, and replies.

In (Hajian et al., 2011), the proposed "Influence Rank" metric implements a modified version of the PageRank algorithm, which is based on the structure and the topology of the network. Specifically, it combines follow-up relationships, mentions, favorites and retweets to identify opinion leaders who are capable of influencing others. The authors conclude that in order to be influential, a user should have influential followers.

In (Jabeur et al., 2012), the authors present a variation of PageRank by introducing two metrics called "InfRank" and "LeadRank", which are based on following, retweeting and mentioning relationships among users. "InfRank" is a variation of PageRank and measures the user influence in terms of his/her ability to spread information and to be retweeted by other influential users. "LeadRank" measures the leadership of a user in terms of his/her ability to stimulate retweets and mentions from other users and especially from other leaders.

Finally, in (Carvalho et al., 2017), the authors present the "MISNIS" framework whose goal is to discover influential Twitter users on a given topic. The framework does so by applying the PageRank algorithm on a graph representing users' mentions found in Portuguese tweets. Moreover sentiment analysis is performed, classifying the messages into three categories namely, positive, neutral, and negative. This work differentiates itself from others in this field in the way that the topics are detected. Instead of performing naive string matching based on the characters of a hashtag, a fuzzy word similarity algorithm is applied utilizing all the contents of a message. Consequently, more relevant tweets on a topic are retrieved despite not containing the exact hashtags or other user indicated keywords.

2.4.1.3. Metrics Based On Machine Learning Techniques

In (Nargundkar and Rao, 2016), social influence is measured by applying the "InfluenceRank" framework. This framework is based on certain features extracted

from profiles (number of tweets, followers, following, member of lists) and tweets over a two-month period. The framework comprises of a regression-based machine learning approach, having "InfluenceRank" as the predictor variable against the set of aforementioned features. Although the work seems promising, the authors claim that, due to the limited number of samples in the training set, the model is not accurate enough.

Another machine learning framework for discovering popular persuasive users is presented in (Fang and Hu, 2016). The authors' persuasiveness metric is pair-wise and is based on three factors: influence, entity similarity, and structural equivalence. Influence depends on the strength of social interactions among users, entity similarity measures how close two profiles are, while structural equivalence measures the structural similarity of two entities according to a distance function. Each of these factors is assigned a probability which denotes the likelihood of persuasion.

The work presented in (Lampos et al., 2014) proposes a framework for predicting user influence by combining textual and non-textual attributes. More specifically, the authors employ the user's basic social information metrics (e.g. the number of followers, followees, mentions and replies), and then by utilizing statistics over the textual data of the tweets, as well as non-linear learning methods and machine learning techniques, a strong prediction performance metric is derived.

Finally, the authors in (Mueller and Stumme, 2017) propose a machine learning methodology for investigating the impact of profile information towards the increase of Twitter accounts' popularity, in terms of their followers' count. Based on the assumption that given names and English words affect the discoverability, profiles were analyzed and categorized into three groups according to the lexical content of the accounts name: i) having a first name, ii) containing English words, or iii) neither of both. The framework consists of three stages to evaluate the popularity dynamics in terms of: a) the content of the name field, b) the profile features, and c) the incorporation of those features in a classifier that identifies the accounts which are likely to increase their popularity. Each group's classifier uses a different model (e.g. Gradient Boosting Machine, Naive Bayes), which is trained with distinct parameters and features, based on the corresponding group. The results showed that the existence of known terms in the name field and the provision of other profile information (e.g. description, profile image, URLs, location) have a strong impact on the number of followers.

2.4.2. Information Flow and Influence

Information flow is vital in all kinds of networks (e.g. social, digital, or computer), and can be affected by the actions or properties of their actors and the sets of dyadic relationships between them. Influential users determine the virality of information and specifically how such information is propagated or diffused. As already mentioned, although propagation and diffusion are often used interchangeably, in this survey we examine them separately. As mentioned before, diffusion defines the spread of information from a starting node towards the rest of the network, while propagation takes into consideration the intermediate nodes as well, which receive, process, and further decide how to handle information. Thus, this topic is divided into two subtopics namely, propagation-oriented and diffusion-oriented. The propagation-oriented approach considers works that employ the propagation of information in

OSNs in order to discover and calculate the impact of influential users whereas the diffusion-oriented approach provides the insights into the identification of influential users being able to boost the diffusion of information in OSNs.

2.4.2.1. Propagation-oriented Approaches

The authors in (Huang et al., 2013) propose an extension of PageRank for measuring influence. They apply their extended PageRank approach on a graph of retweets and user relationships and they consider the social diversity of users and the transmission probabilities of the messages based on the hypothesis that the users inherit influence from their followers. The aim is to explore whether individual characteristics and social actions as well as influence propagation patterns are factors capable of influencing other users.

Similarly, as described in the previous subsection, in (Jabeur et al., 2012) social influence is measured using a variation of PageRank. Specifically, the authors measured the propagation of user influence into the network based on the users' ability to stimulate social actions of others, such as retweets and mentions.

In several cases, the influence metric derived correlates the information propagation with the user's retweeting behavior. Such a study is described in (Romero et al., 2011a), where influence is used for measuring the activity or passivity of Twitter users.

The authors in (Jaitly et al., 2016) propose a methodology to identify influencers in OSNs with the help of online communities which are discovered by applying propagation-based modes. In this case, the structural features (shortest path, closeness, eccentricity, betweeness, and degree) of each node are extracted, while their weighted representation is computed by considering all the features across the network. By using principal component analysis, the most influential nodes are discovered. By applying maximum flow algorithms communities are detected implying a positive attitude towards the influencers.

Social influence and propagation can be used as input in recommendation systems. In (Yuan et al., 2015), influence is considered as a propagated attribute among users in the OSNs. The proposed framework calculates the influence that social relationships have on users' rating behaviors, and incorporates it into recommendation proposals. Two social influence related attributes are considered: user's susceptibility, which is the willingness to be influenced, and friends with high influence.

While the above studies consider the propagation of information towards the discovery of influential users, there are many other works ((Barbieri et al., 2013), (Yang and Leskovec, 2010), (Tang et al., 2016), and (Yi et al., 2016)) that describe frameworks for discovering the propagation of influence in Twitter and its impact on other users.

A framework for modeling the spread of influence in OSNs is developed in (Barbieri et al., 2013). The authors characterize influential users as those generating posts with high probability of being propagated, i.e. retweeted, and simultaneously having a large number of followers. Based on past information cascades influential users are discovered and their social activities and interconnections inside the communities they belong to are analyzed.

In several works, the influence of a node is calculated based on the rate of information spread over the network. For each influenced node an influence function quantifies how many subsequent ones can be affected. This is based on the assumption that the number of newly influenced nodes depends on which other nodes were influenced before. The study described in (Yang and Leskovec, 2010) concludes that the diffusion of information is governed by the influence of individual nodes. Similarly to the previous study (Barbieri et al., 2013), the proposed models are considered as stochastic processes in which information propagates from a node to its neighbors according to a probabilistic rule. The problem lies in discovering influential nodes based on the computation of the expected number of influenced ones (Kimura and Saito, 2006).

A study analyzing the persuasion-driven social influence based on some topic of interest is presented in (Yi et al., 2016). Several influence measurements incorporate the users' social persuasiveness in terms of influence propagation, for quantifying user-to-user influence probability. Based on the same proposed metrics, the framework exploits the topical information, the users' authority and the characteristics of relationships between individuals.

A multi-topic influence propagation model is proposed in (Tang et al., 2016). It is based on user relationships, posts, and social actions. The influence score consists of direct and indirect influence, where the former considers information propagation from retweets by the direct followers, while the latter takes into account the retweets from non-followers. Both of them are related to different topics. The distribution of the users' topics of interest is discovered according to the collected tweets. Then, a topic-dependent algorithm is applied and a multi-topical network is created, in order to identify multi-topic influential users.

A model for demonstrating how social influence can impact the evolution of OSNs by simulating influence propagation and activation processes is proposed in (Yang et al., 2016). In this model, two types of influence namely, locality and popularity, are considered since they have different impact on the network dynamics. Locality affects the information spread through social ties, while popularity has global impact on individuals since it does not rely on network topology.

The Influence maximization (IM) problem is the process of discovering and activating a set of seed nodes to initiate the diffusion process so that the largest number of nodes is reached or influenced. The authors in (Lu et al., 2016) investigate the IM problem and propose a probability-based methodology that enables greedy algorithms to perform efficiently in large-scale social networks in terms of memory and computing costs. The algorithms recursively estimate the influence spread using reachable probabilities from node to node. In (Hung et al., 2016), the authors aim to maximize influence propagation by selecting the most influential intermediate nodes. Therefore, a new optimization problem is formulated which explores the idea of routing multi-hop social influence from the source to a specific target with some time constraint. To achieve this, the topology of the network, the users' influence and the corresponding probability in a specific time frame are taken into consideration. The authors in (Subbian et al., 2016) propose a content-centered model of flow analysis in order to investigate the IM problem. Moreover, the analysis is not based on the users' relationships but on the content of the transmitted messages. The authors apply an algorithm to discover the information flow patterns using content propagation patterns. Then, the influencers are discovered by exploiting those patterns, described

as “boost set selection”, their position and the number of flow paths they participate in. A different approach, on the IM problem is proposed in (Liontis and Pitoura, 2016) and is. The authors claim that it is possible to improve the diffusion process of a subset of the initial seed nodes by using additional resources such as by giving out free samples of a product, engaging in gamification, or other marketing strategies in order to become more influential.

An extension of the IM problem, described as “Influential Node Tracking”, is defined in (Song et al., 2017) where the authors focus on the set of influential nodes dynamically such that the influence spread is maximized at any time. Due to the dynamic nature of the networks, their structure and influence strength associated with the edges constantly change. Consequently, the seed set that maximizes the influence coverage should also be constantly updated. To achieve their goal, the authors compare consecutive snapshots of a network based on the fact that it is unlikely to have drastic changes thereby resulting in great structural similarity.

Finally, there are other works ((Yang and Counts, 2010), (Bakshy et al., 2012), and (Haralabopoulos et al., 2015)) that investigated the discovery of information propagation flows in OSNs. In (Yang and Counts, 2010), a diffusion network is constructed based on user mentions, with constraints on topical similarities in the tweets. The authors claim that given the lack of explicit threading in Twitter, this is the optimal approach of a network to spread information about a specific topic, and that the rate of mentioning of a user is a strong predictor of information propagation. In addition, the authors in (Bakshy et al., 2012) examine information propagation that is related to the exposure to signals about friends' information sharing on Facebook. They found that the users who are aware of that information are significantly more likely to share it faster, compared to those who are not. Although these strong ties are individually more influential, the weak ties, which exceed them in numbers, are responsible for the propagation of information.

2.4.2.2. Diffusion-oriented Approaches

In (Al-garadi et al., 2017), the authors investigate diffusion issues with an improved version of the K-core method (Al-garadi et al., 2016). The authors incorporate a linking and weighting method based on the observation that users’ interactions, namely retweets and mentions, are significant factors for quantifying their spreading capability in a network. In (Zhuang et al., 2017), the authors propose the “SIRank” metric for measuring the users’ spread ability and the identifying influential ones. Initially the users’ spread influence is measured by analyzing the information cascade structure. As each user’s influence is directly related to his/her interaction influence with others, pair-wise metrics are calculated by measuring retweeting contributions, users’ interests and closeness, activity frequency, and retweeting intervals. By quantifying the cascade structure and the user interaction influence on information diffusion, the authors measure the users’ spread influence.

Similarly, the main objective in (Veijalainen et al., 2015) is to investigate the diffusion of messages and the users’ influence, based on the retweet cascade size and its attenuation patterns. The proposed influence measurement depends on the number of users who could potentially get a message either directly or via retweets. The latter affects the proposed cascade size metric and sets the upper limit of users who could potentially see that message. The study concludes that the largest cascades originate

from users with a high number of followers and, the cascade dies out after two or three frequency peaks.

The "retweet" functionality and the retweet counter can be considered as a factor for measuring the "interestingness" of a user's tweets (Naveed et al., 2011a). Based on that, the resulting spread of information is examined in (Kwak et al., 2010). The authors state that the retweet counters are measurements of popularity of the messages and of their authors. According to the study, once a message gets retweeted, it will almost instantly be spread up to four hops away from the source, thus resulting in a fast diffusion after the first retweet. Three different measures of influence, namely the number of followers, PageRank, and the number of retweets, were further compared and evaluated. The results indicated that, in contrast to the third measure, the first two provide similar rankings of influential users, indicating a gap in the influence derived from the number of followers and the popularity of the tweets. Similarly to the results of (Veijalainen et al., 2015), it is observer that the average number of additional recipients is not affected by the number of followers of the tweet source. Thus, the tweet is likely to reach a certain number of audiences via retweets.

A different interpretation of the term "influence" is given in (Bakshy et al., 2011), where the authors relate the user's posting activity (and thus the influence) with the diffusion of the URLs included in posts through retweets. The influence score for a given URL post is calculated by tracking the diffusion of the URL from its source node until the diffusion event is terminated. The work is similar to the one described in (Wei et al., 2016b) where the influence measurement is related to the Influence Diffusion model which provides the influence of a topical spread. However, it differs from (Romero et al., 2011a) in that the diffused influence is studied in terms of activity or passivity of Twitter users solely based on the user's retweeting behavior.

In addition to the point of views discussed above, the following studies involve methodologies for analyzing information diffusion and factors that affect it in OSNs. As already described in the previous subsection, the authors in (Yang and Counts, 2010) claim that, despite the fact that some properties of the tweets predict high information propagation, the users' mention rate is the strongest predictor. The diffusion of information in two social networks, namely Digg and Twitter, is studied in (Lerman and Ghosh, 2010). According to the study, the structure of these networks affects the dynamics of information flow and spread. Information in denser and highly interconnected networks, such as of Digg's, reaches nodes faster compared to sparser networks such as of Twitter's. Due to its structure, information is spread slower, but it continues spreading at the same rate as time passes and penetrates the network further.

In (Bakshy et al., 2012), the information spread is examined regarding the exposure to signals about friends' information sharing on Facebook. The study concludes that social ties greatly affect the users' behavior on re-spreading information in the network. In another work on Facebook, the authors studied diffusion trees of fan pages. The results indicated that there is no solid evidence that a node's maximum diffusion chain length can be predicted (Sun et al., 2009).

The ways in which widely used hashtags spread through interactions among Twitter users are analyzed in (Romero et al., 2011b). Hashtags of different types and topics exhibit different variations of spread. These variations are due to the differences in the spread probability, and to the differences in the extent to which repeated exposures to hashtags continue to affect their diffusion into the network by other users. The authors in (Kafeza et al., 2016) extend their previous work (Kafeza et al., 2014) to identify the

initial set of users who are able to maximize information diffusion. Initially, the users' diffusion patterns are recognized by exploiting their posting activities and history. The proposed algorithm combines them with propagation heuristics in order to achieve the diffusion coverage in the network.

Finally, the authors in (Haralabopoulos et al., 2015) studied the lifespan and the information flows of the Reddit social network based on user-generated content. They were particularly interested in the virality of information and its speed of diffusion in other OSNs. The study concludes that once information is shared within networks, its flow dynamics decrease within the original network.

2.4.3. Network/Graph Properties

The studies presented herein utilize the topology and the structure of the OSNs in order to measure influence or to discover other social dynamics. Usually, in this domain, only a fraction of the whole network is used due to hardware (e.g. RAM, Hard Disk Drive) or complexity limitations.

The framework proposed in (Overbey et al., 2013) aims to automatically identify influential users in topic-based communities. Therefore, a sparser network of Twitter, in terms of the relationships connecting its nodes, is created in comparison to the traditional follower/following network, by leveraging direct communications (mentions and replies). A measure of alpha centrality is employed which incorporates both directionality of network connections and a measure of external importance. As already mentioned in (Jabeur et al., 2012) and (Wei et al., 2016b), influencers are discovered by applying PageRank and newly proposed link-analysis algorithms which are exploiting the topology and the properties of the network, including posting, retweeting and mentioning relationships among users. In (Almgren and Lee, 2016) influence is measured by applying a hybrid framework that integrates both users' structural location and attributes. A user's location is found by applying several centrality analysis algorithms (in-degree, weighted in-degree, eigenvector, and PageRank) while the attributes (i.e. activeness) are measured by adapting the contribution measurement, which is used by Flickr, and is based on the number of uploaded photos.

The authors in (Jaitly et al., 2016) propose a methodology for the identification of influencers by exploiting structural features. Specifically, the shortest path, closeness, eccentricity, betweeness centrality, and degree of each node are extracted and their weighted representation is computed by considering all the features across the network. The most influential nodes are discovered by using principal component analysis. Moreover, by applying maximum flow algorithms communities are detected. The identified communities imply a positive attitude towards the influencers.

In (Yang et al., 2016), the authors propose a framework to demonstrate how social influence can impact the evolution of OSNs by simulating influence propagation and activation processes. In this framework, two types of influence are introduced, that have different effects on the network dynamics. The first type is "locality" which affects information diffusion through social ties, while the second is "popularity" that does not rely on network topology but has a global impact on individuals.

All the above studies try to identify influencers according to the information derived in particular periods of time, similar to a compilation of different and static sequences.

Below, we analyze other works where related issues are considered under properties and concepts that belong to dynamically evolving and complex networks.

A dynamic index data structure for influence analysis on an evolving network is presented in (Ohsaka et al., 2016). The indexing method is able to recognize and incorporate all the graph updates in order to efficiently answer the queries on influence estimation and maximization on the latest graph edition. Several optimized techniques (e.g. a reachability-tree-based technique for edge/vertex deletions, a skipping method for vertex additions, and a counter-based random number generator for the space efficiency) are incorporated to reduce time and space requirements.

In (Song et al., 2017), the Influential Node Tracking problem is defined as an extension of the Influence Maximization in dynamically evolving networks. Due to their nature, the structure and influence strength associated with the edges change constantly. Therefore, the authors consider the dynamic network as a set of static ones, and compare consecutive snapshots under the assumption that it is unlikely to have drastic structural changes.

Another work in this area is presented in (Yang et al., 2017). A dynamic network is modeled as a stream of edge weight updates. Under the assumptions of the linear threshold model, two versions of the problem are considered: the discovery of nodes having influence greater than a specified threshold, and finding the top-k most influential nodes. The proposed algorithm incrementally updates the sample random paths against network changes by considering efficiency in both space and time usage.

Apart from discovering influential users, the topological and structural attributes of the networks can be used towards the context-based identification of users' interests and similarities. For example, a community detection in OSNs approach is proposed in (AlFalahi et al., 2013) using node similarity techniques. A virtual network is created, where virtual edges are inserted based on the similarity of the nodes in the original network. The similarity is calculated using the Jaccard Measure. The proposed algorithm is then applied on the generated virtual network.

Similarly, in (Karidi, 2016) proposes a semantic followee recommender system in Twitter which exploits users' tweets in order to build their interest profiles. An interest graph is created by using specific semantic knowledge graphs that contain a variety of topics. These topics are then mapped and suggested to the users. User interest metrics are calculated using graph theory algorithms such as the Steiner Tree and the "InterSim" (Interest Similarity) ones.

Another context-oriented approach is presented in (Kalloubi et al., 2016), where the context of Twitter posts is retrieved using the DBpedia knowledge base and graph-based centrality theory. A graph of contextualized and weighted entities for each tweet is constructed, and two types of similarity metrics are introduced. The "local" similarity measures the proximity of two entities in terms of the context in which they occur. When a user request is made, the "global" similarity is calculated from this request and from the available tweets.

2.4.4. Applications

In this section, we consider the influence metrics presented in Section 2.4.1, to present research efforts that provide solutions for opinion makers, data analysts and

information scientists as services or applications. This topic is further divided into three subtopics namely a) rank-oriented, b) recommendation-oriented and c) other application domains, such as sentiment analysis, and event detection, .

2.4.4.1. Ranking

To rank OSN users according to specific social attributes, the work described in (Kong and Feng, 2011) presents a qualitative measurement of tweets that determines the influence of their authors in order to present a tweet-centric topic-specific author ranking. The quality of the tweet is evaluated according to the topic focus degree, the retweeting behavior, and the topic-specific influence of the users who retweeted this topic. In (Li et al., 2013b), the authors propose an influence ranking method under the assumption that the user influence is based on the followers' influence and their interactions. The authors found that user A can exert more influence over user B if user A posts tweets strongly related to user B. The proposed measurement is a variation of the PageRank and is based on a similarity factor between published tweets, on a graph of following, retweets, mentions and replies.

The authors in (Wei et al., 2016a) propose a framework for discovering topic-specific experts in Twitter by employing two distinct metrics. First, the users' global authority (influence) on a given topic is calculated offline by exploiting three types of relations (i.e. follower relation, user-list relation, and list-list relation). Second, the similarity between the users' generated tweets and that topic is computed online. By leveraging the users' topical influence and similarity, those who have the highest-ranking scores are regarded as experts in that domain.

The problem of topic-sensitive opinion leaders' identification in online review communities is also investigated in (Miao et al., 2016), where a two-staged approach is presented. Initially, the opinion leaders' expertise and interests are derived from their tags found at the description of the products. Then, a computational approach measures the leaders' influence and ranks them according to not only the link structure of customer networks, but also according to their expertise and interests. The influence depends on the topical similarity between reviewers on a specific topic.

The authors in (Francalanci and Husain, 2017) created "NavigTweet", an influence-based visualization framework to explore Twitter followers relationships, by browsing the friends' followers network and to identify key influencers based on the actual influence of the disseminated content. The top influencers are identified by both user-level (i.e. followers, following, tweets, lists) and content-based (hashtags, URLs, retweets, favorites, mentions) parameters. Then, based on the above, the "Analytical Hierarchy Process" is used to rank Twitter users.

An influence learning-based recommender is presented in (Chen et al., 2016b) for making suggestions to informative users whose posts are highly associated with those of the target users. Ranking learning techniques are designed to analyze user behavior and to model their preferences based on their social interactions (e.g. replies, likes). Moreover, the social influence among users is incorporated into the learning model to enhance the learned preferences. In another application described also in Section 2.4.2.2 the authors propose the "SIRank" metric for measuring the users' spread ability and for identifying the influential ones (Zhuang et al., 2017).

2.4.4.2. Recommendation

The various studies presented in this section describe approaches on recommendation systems which utilize the available information in OSNs for proposing social content or accounts based on the users' profiles. An interesting problem in the area of social network recommendation systems is to define a set of similar users to follow.

The friend recommendation problem in Flickr is studied in (Huang et al., 2016), mainly from the viewpoint of network correlation. The authors assume the hypothesis that each user has many different social roles in OSNs. For each role different social sub-networks are formed, which are aligned in order for the correlations among them to be found through a weighted tag feature selection. When recommendations are made, the similarities of the tag features, among the new and the existing users, are calculated. The more similar the tags are, the more users there are who are similar in terms of those tags.

A semantic followee recommender system in Twitter is proposed in (Deb et al., 2016). This system integrates content-based filtering approaches based on tweet analytics, and popularity identification among users using collaborative-filtering over the friendship network, along with publicly available knowledge resources (i.e. Wikipedia, WordNet, Google corpus). The aim is to classify the tweets into six classes and to label the users as a recommendation service. The application of the Kalman filter enables noise removal and the prediction of future tweet patterns leading to the new multi-labeling of the users.

Similarly, the work described in (Karidi, 2016) exploits the users' tweets for building their interest profiles and for producing recommendations over a semantic knowledge graph that contains a variety of topics. Using graph theory algorithms (as explained in Section 2.4.3), the authors can recommend similar users. A ranking-based followee recommendation scheme in microblogging systems that is based on the latent factor model is proposed in (Chen et al., 2016a). To model user preferences both tweet content (original posts and retweets) and social relation information (followers, followees) are taken into consideration. Another followee recommendation methodology that builds interest profiles is proposed in (Hannon et al., 2010). These profiles are built by exploiting not only the user generated content but also the content of their directly related ones (followers, followees).

A framework for discovering similar accounts in Twitter based only on the "List" feature is proposed in (Kanungsukkasem and Leelanupab, 2016). This functionality allows the users to create their own lists by adding any account they wish. The authors claim that this feature is considered a form of crowd-sourcing. The hypothesis of the methodology is that when two accounts are contained in the same list they should be similar or related to each other. Therefore, the proposed measurement relies on the number of lists that a specified account and a potentially similar one are listed together.

In (Ma et al., 2011) a matrix factorization framework with social regularization is proposed for improving recommender systems by incorporating social network information. Social regularization includes two models for representing social constraints, and those methods are based on users-friends similarity at an individual and average level. Each social link is then weighted in accordance with the similarity among the users, allowing the exploitation of friends based on the rating similarity.

As already described in the previous subsection, ranking learning techniques are designed to provide recommendations based on the analysis of user behavior, preferences, and social interactions (Chen et al., 2016b). In addition, as mentioned earlier in Section 2.4.2.1, we can use related attributes being propagated through the social network because the effects of friends who have strong influence or are subject to be influenced by a user are highly related to the recommendation processes (Yuan et al., 2015).

The recommendation system proposed in (Li et al., 2017) is based on the users' personal interests. In fact, explicit social features such as the users' topic-level influence, topic information, and relations are incorporated into a framework for improving recommendation results. Two kinds of influence are introduced: direct, which is identified by studying the communication records between users, and indirect, which is identified by applying the social status theory for the discovery of latent relationships. In both cases, positive and negative influences are also identified. Moreover, topic information is added into the structural analysis of indirect influence. A distributed learning supervised algorithm is applied which takes into consideration the aforementioned influence measurements and provides the users' forwarding behaviors, which can be leveraged to provide improved recommendations.

Considerable attention has also been paid to recommendation systems for suggesting personalized streams of information ((Phelan et al., 2011), (Chen et al., 2010), (Tang et al., 2016), and (Zhang et al., 2016)).

"Buzzer" is such kind of a service for proposing news articles to Twitter users, by not only mining terms from their timeline, but also from their friends' timelines (Phelan et al., 2011). These terms act as ratings for promoting and filtering news content. The methodology described in (Zhang et al., 2016) is based on the same principles but it also incorporates additional factors affecting the interest of a user on a tweet, such as its quality, the number of retweets, and the importance of its publisher.

URLs as a recommendation factor in Twitter are studied in (Chen et al., 2010) in terms of directing the users' attention in more focused information streams, namely to Twitter posts, from the viewpoint of personalized content suggestion. The authors explored three separate dimensions in designing such a recommender: the sources of the URLs, the users' area of interest, and social information.

The authors in (Tang et al., 2016) propose a multi-topic influence diffusion model based on user relationships, posts, and social actions. The influence score consists of direct and indirect influence. The first is determined by information propagation (retweets) by the direct followers. The latter depends on the retweets from non-followers. Both of them are related to different topics. Based on the users' collected tweets, the distributions of their topics of interest are found along with their generation probability. Finally, a multi-topical network is created to which a topic-dependent algorithm is applied in order to identify the multi-topic influential users while the most influential user will be used during the recommendation process.

Finally, recommenders can also be used for suggesting items on users. The work described in (Zhang et al., 2016) is based on the observation that a user's purchase behavior is influenced by both global and local influential nodes which in turn define implicit and explicit social relationships respectively. Therefore, a dual social influence framework formulates the global and local influence scores as regularization

terms, and incorporates them into a matrix factorization-based recommendation model.

2.4.4.3. Other Application Domains

The studies presented in this section exploit user' influence in OSNs in other domains (such as sentiment analysis, user polarity and event detection) than those described in Sections 2.4.4.1 and 2.4.4.2.

Moreover, the identification of influential users in Twitter is based on a combination of the users' position in the networks derived from the Twitter relations, the sentiment of their opinions, and the textual quality of their tweets. Thus, in (Bigonha et al., 2012), the authors propose a centrality measure that combines betweeness and eigenvector centralities, in-degree and the follower-followee ratio on graphs of relationships, mentions, replies and retweets. Using sentiment analysis techniques, the users are classified into those having positive, negative or neutral tweets.

Another sentiment-based framework is proposed in (Zhao et al., 2014), where sentiment is discovered through exchanged messages among users in online health communities. The metric focuses on the sentimental effect of inter-personal influence on individuals and reflects a user's ability to directly influence other users' sentiments.

The authors in (Piškorec et al., 2016) investigate whether the users' friendship network can interfere with the peer and the external influences. The experiment takes place during an on-line voting procedure in Facebook. The analysis of the users' demographics and votes showed a strong homophily among the communities and the friends' votes. The authors analyzed both peer and external influences in order to explain the activation of voters. Peer influence propagates from recently activated friends while external influence from news agencies affects all users uniformly.

Finally, a story-tracking framework based on hashtags in OSNs is proposed in (Poghosyan and Ifrim, 2016). The storyline extraction is modeled as a pattern mining and real-time retrieval problem. The most popular news stories, which have been assigned hashtags, are detected by mining frequent hashtag pattern sets. Using query expansion on the original hashtags new story articles are retrieved. The pattern set structure enables hierarchical and multiple-linkage representation of the news.

2.4.5. Comparison of Related Works

In order to provide comparative insights from the above reviewed articles that refer to online social influence, we provide Table 2.3 found in Appendix A: Comparison of Reviewed Articles. For each reviewed article, the first three columns denote its category according to our classification scheme (see Figure 2.1), the section number, as well as its reference. It should be noted that in many cases, a study does not fall within the scope of only one topic, thus demonstrating that research efforts are strongly related to each other. It should be noted that in many cases, a study does not fall within the scope of only one topic, thus demonstrating that research efforts are strongly related to each other. In the rest columns, we place a mark of "Yes" (✓) or "No" (✗), to indicate whether the studies employ or propose metrics and characteristics based on:

a. Relationship: followers (Fs) and followees (Fing),
b. Behavioral/Conversational activities: posts (P), re-posts (RP), favorites/likes (FL), mentions (M), replies (R), and
c. Domain/Content analysis: works that are applied on specific topics (T) or works that require content analysis (CA).

2.5. Online Social Semantics

In this section, we study the semantics and their role as the second major aspect of OSNs. Specifically, we analyze related works based on Semantic Web technologies along with network theory and graph properties for transforming unstructured data into Linked Data, topic identification, detection of similar users and communities, as well as user personalization (e.g. interests, suggestions, and so on).

2.5.1. Social Modeling

As the adoption of semantics and Linked Data increases, a large number of works have emerged covering aspects of semantic modeling in OSNs. In this section, we present approaches which adopt semantics for modeling the logical topology and structure of social networks and media as well as the information they disseminate.

One of the first studies in this domain is (Hepp, 2010), where the use of a specific syntax is proposed for creating a common knowledge representation, by incorporating RDF-like syntaxes into Twitter posts. The use of such statements enables users to freely define relations such as hierarchical or equality relations among hashtags. Hence, an ontology of hashtags is collaboratively created which can be exploited for the resolution of synonymous hashtags or other simple reasoning tasks.

The authors in (Celik et al., 2011) propose a framework for enriching Twitter messages with semantics relationships by analyzing Twitter posts. These relationships are identified among persons, products, and events and that are utilized in order to provide query suggestion to the users.

Another work on the enrichment of Twitter messages with semantics is described in (Abel et al., 2011). The authors attempt to create user profiles by exploiting Twitter posts using Semantic Web technologies. In order to capture the users' interests, the URLs of news articles found in tweets are utilized. Lexical analysis is applied on their content so that the relationships between the entities in news articles (representing the interests) can be discovered. These entities are then semantically related to those tweets.

Social semantics can be exploited in the development of semantic recommender systems. The studies (Deb et al., 2016) and (Karidi, 2016), which were analytically presented in Section 2.4.4.2, propose two semantic followee recommender systems for Twitter. Their aim is to build user interest profiles by exploiting the users' posted messages (Deb et al., 2016), (Karidi, 2016), friendship network (Deb et al., 2016) and publicly available knowledge bases (i.e. Wikipedia, WordNet, Google) (Deb et al., 2016), which are then used during the recommendation process.

A framework for inferring user interests in Twitter is also proposed in (Besel et al., 2016). In contrast to the ones described above, this framework is based on the users' followees and the content they consume, rather than on their original posts. The proposal is based on the hypothesis that famous people maintain accounts that are being followed by a large number of users. The Wikipedia articles of the former are discovered, linking to a higher level of categories and hierarchies, which become an implicit expression of the users' interests.

The methodology presented in (Xia and Bu, 2012) is capable of handling large-scale networks and of generating weighted semantic ones. These networks are created by using comments from a Chinese social network. The methodology focuses on the "giant component" of the derived network in order to reduce the computational complexity so that larger networks can be better handled.

The work described in (Packer et al., 2012) associates tweets with a given event by utilizing the structured information found in them. The initial pool of terms for the retrieval of the messages is manually provided. The final associations take place by applying query expansion techniques and by utilizing the relationships derived by the semantified data.

The authors in (Wang et al., 2016) create graphs of hashtags found in tweets and utilize their relational information in order to discover latent word semantic connections in cases where words do not co-occur within a specific tweet. Sparseness and noise in tweets are handled by exploiting two types of hashtag relationships: i) explicit ones which refer to hashtags that are contained in a tweet, and ii) potential ones which refer to hashtags that do not appear in a tweet but co-occur with others. Finally, the hashtags and words which have the highest probability to appear on a specific topic, are discovered.

The following studies employ Semantic Web technologies, ontologies and the DBpedia knowledge base, which is a semantified version of Wikipedia, to achieve their goals.

The study in (Shinavier, 2010) introduces a semantic data aggregator in Twitter, which combines a collection of compact formats for structured microblog content with Semantic Web vocabularies. Its purpose is to provide user-driven Linked Data. The main focus of this work is on posts and specifically on their creators, content and associated metadata.

Another framework which utilizes semantic technologies, common vocabularies and Linked Data in order to extract microblogging data from scientific events from Twitter, is proposed in (Vocht et al., 2011). In this work, the authors attempt to identify persons and organization related to them based on geospatial and topic entities.

The authors in (Slabbekoorn et al., 2016) propose an ontology-assisted topic modeling technique for determining the topical similarities among Twitter users. The entities found at the posts are mapped to classes of the DBpedia ontology, using the DBpedia Spotlight tool, and are used for the labeling of clusters. Moreover, the topical similarities among individuals on different topics are calculated using ranking techniques, which define the structure of the resulting graphs. Based on these graphs, a quasi-clique community detection algorithm is applied for the discovery of topic clusters, without predefining their target number.

Another work using the DBpedia knowledge base is (Kalloubi et al., 2016), where a framework is proposed for retrieving the context of posts in Twitter by applying the graph-based centrality theory. Entities from tweets, in terms of words, are extracted and related to DBpedia URIs in order for semantic concepts to be discovered. Based on the graph centrality theory, a graph of contextualized and weighted entities for each tweet is constructed.

2.5.2. Social Matching

The studies presented in this section exploit the use of social semantics for identifying similar properties and activities with respect to user-generated content, description of real-life events, as well as for revealing user interests and behavioral patterns across different online social media users. Thus, we divide this topic into two subtopics, namely a) User-oriented (e.g. similar user recommendation, user preferences), and b) Topic and Event-oriented (e.g. topic profiling and user interest, event detection, product marketing and others).

2.5.2.1. User-oriented Matching

Despite the fact that the set of social semantics of each account in an OSN is unique as they depend on personal social activities, common patterns among them can be recognized. These patterns can be exploited to enable the discovery of the users' social behavior and preferences.

The study presented in (Räbiger and Spiliopoulou, 2015) describes a framework using supervised learning for distinguishing users in OSNs according to their influence and reveals the communities they belong to. The authors do not define a new influence measure, but discovered predictive properties associated with the users' activity level and involvement in those communities. The supervised learning is based on follow-up relationships, interactions (mentions, replies), the structure and activity of the network, the centrality of users, and the quality of the tweets. The study concludes that these relationships are the most important ones for identifying influential users.

The aim in (Sun and Ng, 2012) is also the identification of influential users based on their interactions in their posts on a given topic. Toward this end, a graph model representing the relationships of the posts is created, which is then transformed into a user graph. The latter is used for the discovery of influential users by considering properties and measures from both graphs. Similarly, as described in Section 2.4.1.2, the authors in (Carvalho et al., 2017) follow the same approach and apply the PageRank algorithm on that graph for the detection of influencers. The posts belonging to a specific topic are discovered through a fuzzy word similarity algorithm which utilizes all the contents of the messages.

In (Kong and Feng, 2011)influencers are regarded as those generating tweets of high quality. Their quality is evaluated according to a set of parameters such as the topic focus degree, the retweeting behavior, and the topic-specific influence of the users who retweeted those messages.

The framework proposed in (Overbey et al., 2013), as presented in Section 2.4.3, aims identify influential users in topic-based communities. A measure of alpha centrality is

employed on a graph derived from direct communications, which incorporates both directionality of network connections and a measure of external importance.

Influential users are discovered in (Weng et al., 2010) by applying as an extension of the PageRank algorithm, which takes into consideration both the topical similarity among users and their link structure. It is claimed that due to homophily, that is the tendency of individuals to associate and bond with others having similar interests, most of the "Follower-Following" relations appear. This work also suggests that the active users are not necessarily influential.

In Section 2.4.4.1 we described a study (Miao et al., 2016) where the problem of topic-sensitive opinion leaders' identification in online review communities is investigated. Toward this end a two-staged approach is presented. Initially the opinion leaders' expertise and interests are derived from their tags found at the description of the products. During the next stage a computational approach measures the leaders' influence and ranks them according to not only the link structure of customer networks but also to their expertise and interests. The influence depends on the topical similarity among reviewers on a specific topic.

The task of the topic experts' identification, namely influential users on specific domains, is also presented in (Liu et al., 2014). A post-feature based approach is proposed which utilizes nine kinds of features reflecting how the users interact. Their aggregation results in the production of three different kinds of influence measurements.

The authors in (Ramírez-de-la Rosa et al., 2014) claim that a user's influential level can be detected by considering the writing style and behavior within the OSNs. Therefore, they proposed 23 features of user profiles (e.g. presence of hashtags, URLs, self-mentions, number of followers and tweets) and 9 features of tweets (e.g. extension, frequency, quality, number of retweets) that can affect influence impact. By applying machine learning algorithms, the most influential users are identified.

A framework exploiting machine learning techniques for discovering top persuasive users in OSNs is described in (Fang and Hu, 2016). The proposed persuasiveness metric is pair-wise and is based on three factors: influence, entity similarity, and structural equivalence. Influence depends on the strength of those social interactions between users. Entity similarity measures how close two profiles are. Structural equivalence measures the structural similarity of two entities according to a distance function. Each of these factors is assigned a probability which denotes the likelihood of persuasion. A machine-learning algorithm is used to predict these probabilities.

The rebroadcasting behavior of users in OSNs is studied in (Zhang et al., 2017).A model is proposed which examines three aspects: the role of content, the content-user fit, and the social influence. The "content-user fit" measure considers the interaction between the message content and user interests. As in (Hassan et al., 2016), influence measures the susceptibility of users for identifying those whose posts affect the reposting behavior of others. In order to discover the users' interests, the well-known Latent Dirichlet Allocation (LDA) (Blei et al., 2003) methodology is applied on each message. The study concludes that the rebroadcasting of messages does not depend only on its content but also on its relevance to a user.

The LDA topic modeling approach is also applied in (Nigam et al., 2016), where a user centric topic discovery framework is proposed. The users' tweets are analyzed for identifying their interests and for creating personalized topic profiles. Toward this

end a Part-Of-Speech (POS) tagger extracts the nouns of the tweets, which are provided to a search engine to retrieve the top documents based on their relevance. Using LDA on the content of these web pages the final topics are provided.

Another framework, which was also described in Section 2.5.1, is employed for inferring user interests in Twitter (Besel et al., 2016). Contrary to the previous frameworks, this one is based on the users' followees and the content they consume, rather on than their original posts. The proposal is based on the hypothesis that famous people maintain accounts being followed by a large number of users. The Wikipedia articles of the former are discovered, linking to a higher level of categories and hierarchies, which become an implicit expression of the users' interests.

The following studies exploit the social semantics in OSNs in order to propose query expansion techniques for providing an enriched coverage of information needs.

The study in (Zhou et al., 2012) describes a query expansion framework that takes into account the users' preferences which are derived by analyzing microblog posts and hashtags related to the targeted users.

Another query expansion approach is proposed in (Reda et al., 2011). It takes into consideration the similarity between tags composing a query and the social proximity between the query and the user's profile. Its aim is to assist users by refining and formulating their queries and by providing them with information relevant to their interests.

The research effort of (Packer et al., 2012) we have presented in Section 2.5.1, attempts to associate tweets with a given event, by utilizing their structured information. The application of query expansion techniques and the relationships derived from the semantified data result in those associations.

The story-tracking framework of study (Poghosyan and Ifrim, 2016) is modeled as a pattern mining and real-time retrieval problem. The most popular news stories, assigned with hashtags, are detected by mining frequent hashtag pattern sets. Using query expansion on the original hashtags new story articles are retrieved. The pattern set structure enables hierarchical and multiple-linkage representation of the articles.

The authors in (Efron, 2010) attempt to identify several hashtags relevant to a given query, that can be used to expand it thus leading to more accurate content retrieval. The proposed method leverages statistical techniques to build probabilistic language models for each available hashtag through a suitable microblog posts corpus.

Several studies ((Celik et al., 2011), (Abel et al., 2011), (Vocht et al., 2011), and (Slabbekoorn et al., 2016)), which were also presented in Section 2.5.1, utilize semantic technologies and related protocols to provide expanded query suggestions or to represent user preferences and similarities. The authors of (Celik et al., 2011) propose a framework for enriching Twitter messages with semantic relationships by analyzing Twitter posts. These relationships are identified among persons, products, and events and are utilized in order to provide query suggestions to the users. The authors attempt to identify persons and organizations related to them based on geospatial and topic entities. The study in (Blei et al., 2003) uses Semantic Web technologies for the creation of user profiles by analyzing Twitter posts. In order to capture the users' interests, the URLs of news articles found in tweets are used. A lexical analysis is applied on their content in order to discover the relationships between the entities in news articles (representing the interests) which are then semantically related to those tweets. The framework in (Vocht et al., 2011) exploits

common vocabularies and Linked Data in order to extract microblogging data regarding scientific events from Twitter. Finally, an ontology-assisted topic modeling technique for determining the topical similarities among Twitter users is proposed in (Slabbekoorn et al., 2016). The entities found at the posts are mapped to classes of the DBpedia ontology and are used for the labeling of clusters. Moreover, the topical similarities among individuals on different topics are calculated using ranking techniques, which define the structure of the resulting graphs. Based on these graphs, a quasi-clique community detection algorithm is applied for the discovery of topic clusters without predefining their target number.

In Section 2.4.3 and Section 2.4.4.2, we presented several studies studies ((Ma et al., 2011), (Huang et al., 2016), (Karidi, 2016), and (Kanungsukkasem and Leelanupab, 2016)) that attempt to adequately describe user characteristics, in order to discover similarities among them. In (Ma et al., 2011) a matrix factorization framework with social regularization is proposed for improving recommender systems by incorporating social network information. Each social link is weighted based on the similarity among the users, allowing the exploitation of friends differently according to the rating similarity.

The friend recommendation problem in Flickr is studied in (Huang et al., 2016), from the viewpoint of network correlation. The authors assume that each user has many different social roles in OSNs. During each role different social sub-networks are formed which are aligned in order to find the correlations among them through a weighted tag feature selection. When recommendations are made, the similarities of the tag features among the new and the existing users are calculated. The more similar the tags are, the closer the users should be.

The author in (Karidi, 2016) proposes a semantic follower recommender system in Twitter which exploits the users' tweets in order to build interest profiles. An interest graph is created using specific semantic knowledge graphs containing a variety of topics, which are then mapped to the users according to their semantic relevance to the topics. Using graph theory algorithms the user interest similarity is calculated which is used during the recommendation process.

A framework for discovering similar accounts in Twitter based only on the "List" feature is proposed in (Kanungsukkasem and Leelanupab, 2016). This functionality allows the users to create their own lists by adding any account they wish. The authors claim that this feature is considered a form of crowd-sourcing. The hypothesis of the methodology is that when two accounts are present in the same list they should be similar or related to each other. Therefore, the proposed measurement relies on the number of lists that a specified account and a potentially similar one are listed together.

2.5.2.2. Topic and Event-Oriented Matching

As we have already mentioned, social semantics patterns can be used to identify user interests or topics of discussion such as real-life events.

The studies described in (Pal and Counts, 2011), (Sun and Ng, 2012), (Kong and Feng, 2011), and (Yi et al., 2016) are specialized in discovering the most influential authors in Twitter on a specific topic. In (Pal and Counts, 2011), the authors suggest a set of metrics based on original tweets, replies, retweets, mentions and friendship

relationships. In (Sun and Ng, 2012), which was also described in Section 2.5.2.1, these metrics are discovered by considering properties and measures on user-post graphs, while in (Kong and Feng, 2011), presented in Section 2.4.4.1, influencers are regarded as those generating tweets of high quality. A different kind of social influence, a persuasive one, is proposed in (Yi et al., 2016). The proposed measurement depends on topical information, the users' authority and the characteristics of relationships among individuals.

The authors in (Subbian et al., 2016) propose a content-centered model of flow analysis for investigating the Influence Maximization problem on topic-specific influencers. As also described in Section 2.4.2.1, this analysis is not based on the users' relationships, but on the content of the transmitted messages. Influencers are discovered by exploiting information flow patterns, and their position, as well as the number of flow paths they participate in.

A framework for determining the relevance of Twitter messages for a given topic is introduced in (Tao et al., 2012). Two feature categories are identified, i.e., features related to the user query and, thus, calculated as soon as the latter is formed, and features that are not related to this query but are inherent posts and are therefore calculated when they are modified.

In Section 2.5.1, we presented an approach that associates tweets with a given event, by utilizing their structured information (Packer et al., 2012). The application of query expansion techniques along with the relationships deriving from the semantified data, result in those associations.

Another topic-oriented framework for Twitter is presented in (Michelson and Macskassy, 2010). Its aim is to discover the users' topics of interest by examining the entities found in their posts, which may be mentions or plain text (in OSNs the mentions are words prefixed with "@"). The Wikipedia knowledge base is leveraged in order to disambiguate those entities and the topics of interest to be defined (e.g. the term "apple" may refer to the fruit or to the multinational technology company).

The work in (Nigam et al., 2016), also described in Section 2.5.2.1, presents an LDA (Blei et al., 2003) topic profile modeling approach for the discovery of the users' interests. A POS tagger extracts the nouns from their tweets, which are then provided to a search engine to retrieve the top related web pages. LDA on the content of these web pages is used to discover the final topics.

Topic profiling using the Wikipedia knowledge base is also studied in (Lim and Datta, 2013). The topics are discovered based on the posted hashtags of the Twitter accounts and of their friendship relationships. The celebrities (accounts of popular people) who are followed by those accounts are the primary source for discovering users' interests. Those interests derive from a classification based on Wikipedia. The indicators along with the posted hashtags infer the topics of interest of the accounts.

Similarly, the work in (Kapanipathi et al., 2014) focused on extracting the interests of Twitter accounts based on their generated messages. The proposed methodology leverages the hierarchical relationships found in Wikipedia in order to infer user interests. The authors claim that the hierarchical structures can improve existing systems to become more personalized based on broader and higher level concepts (e.g. the concept "Basketball" is more generic that the term "NBA").

The authors in (Slabbekoorn et al., 2016) propose an ontology-assisted topic modeling technique for determining the topical similarities among Twitter users. The entities

found in the posts are mapped to classes of the DBpedia ontology and are used to label clusters. Moreover, the topical similarities between individuals on different topics are calculated using ranking techniques, which define the structure of the resulting graphs. Based on the graphs, a quasi-clique community detection algorithm is applied for discovering topic clusters without predefining their target number.

Another topic-oriented framework, also presented in Section 2.4.3, which uses the DBpedia knowledge base is proposed in (Kalloubi et al., 2016). The context of Twitter is mapped to DBpedia entities and the graph-based centrality theory is applied for assigning weights to the entities of the examined messages.

Topic profiling is also exploited in recommendation systems. Such studies include (Deb et al., 2016), (Karidi, 2016), and (Li et al., 2017), which were also presented in Section 2.4.4.2. The first two ((Deb et al., 2016) and (Karidi, 2016)) describe methodologies for the creation of semantic followee recommender systems for Twitter. These studies are based on the classification of the content of tweets and the users that generated them and on semantic knowledge graphs containing a variety of topics being mapped to users respectively. The recommender described in (Li et al., 2017) discovers personal interests by applying a distributed learning supervised algorithm by taking into consideration explicit social features such as the users' topic-level influence, topic information, and social relations.

A framework for discovering topic-specific experts in Twitter by employing two distinct metrics is presented in (Wei et al., 2016a). The first metric measures the users' global authority on a given topic, while the other metric provides the similarity between the users' generated tweets and that topic. By leveraging the topical influence and similarity, the users who have the highest-ranking scores are regarded as experts in that domain.

In Section 2.4.2.1, we described a multi-topic influence propagation model based on user relationships, posts and social actions (Tang et al., 2016). The influence score consists of direct and indirect influences, related to different topics. The distribution of the users' topics of interest depends on the content of the disseminated messages. The proposed topic-dependent algorithm is applied and a multi-topical network is created in order to identify multi-topic influential users.

Another framework exploiting both user interests and social influence is presented in (Zhang et al., 2017). The rebroadcasting behavior of users in OSNs is studied in this work. The proposed model examines three aspects; the role of content, content-user fit, and social influence. The "content-user fit" measure considers the interaction between the message content and the user interests. Study (Räbiger and Spiliopoulou, 2015) presented a framework using supervised learning for discovering the communities the users belong to and for identifying the most influential ones.

2.5.3. Community Detection

Community detection is not only useful for the analysis of OSNs but also for understanding the structure and the properties of complex networks. The aim is to group their nodes into potentially overlapping sets that share common attributes and characteristics. The following studies propose various approaches to detect communities in OSNs.

In contrast to the traditional techniques, the approach in (Yang et al., 2013) uses both the structure of the network and the attributes of the nodes in order to develop an algorithm for detecting overlapping communities.

Another algorithm for detecting communities is presented in (Qi et al., 2012). This algorithm is based on the content of the edges derived from the users' pair wise interactions. According to the authors, this algorithm provides richer insights into the communities because it depicts the nature of the interactions more effectively.

As described in Section 2.5.1, the methodology in (Xia and Bu, 2012) aims at discovering the latent communities in large-scale networks and generates weighted semantic ones. The latter are created using the information extracted from users' comment content.

The approach in (AlFalahi et al., 2013) presented also in Section 2.5.1, detects communities using node similarity techniques. A virtual network is created where virtual edges are inserted into the original network based on the similarity of the nodes. This similarity is calculated using the Jaccard Measure. The proposed algorithm is then applied on the generated virtual network.

(Slabbekoorn et al., 2016) propose a topic modeling technique among Twitter users using the DBpedia ontology. A community detection algorithm is applied on the users' graph to discover the topics.

Community detection can also be used for the identification of influential users in OSNs. The work in (Jaitly et al., 2016) proposes such a methodology by applying maximum flow algorithms on a weighted representation of the network by considering structural features such as shortest path, betweeness, closeness and degree centralities.

Another framework, also presented in Section 2.5.2.1, for discovering top persuasive users in OSNs is described in (Fang and Hu, 2016). This framework is based on machine learning techniques and depends on three factors: influence, entity similarity, and structural equivalence. Each of these factors is assigned a probability (denoting the likelihood of persuasion) that has been derived by using a machine-learning algorithm. Finally, another study that uses communities to identify influential users is presented in (Barbieri et al., 2013). The authors analyzed the social activity and the interconnections of the users inside the communities they belong to. The communities are utilized to develop a framework for modeling the spread of influence by identifying the most influential users.

2.5.4. Comparison of Related Works

We provide here comparative insights (Table 2.4 in Appendix A: Comparison of Reviewed Articles) from the above reviewed articles that refer to online social semantics. For each reviewed article, the first three columns denote its category according to our classification scheme (see Figure 2.1), the section number, as well as its reference. In the other columns, we place a mark of "Yes" (✓) or "No" (✗), to indicate whether the studies employ or propose metrics and characteristics based on:

a. Network Structure (NS): includes social follow-up relationships or other types of network linking (e.g. based on mention, reply actions).

b. Behavioral/Conversational activities: posts (P), re-posts (RP), favorites/likes (FL), mentions (M), replies (R), Contextual analysis: works for building user profiles or providing personalized information (P), and those applied on specific topics (T).

c. Use of Knowledge Bases (KB): works that use publicly available, open or crowd-sourcing-based resources (e.g. Wikipedia, DBpedia).

d. Use of semantics: modeling unstructured data using the RDF protocol, use of (existing or new) ontologies (O).

e. OSN Entities: hashtags (H) and web URLs distributed in social content.

2.6. Modeling the quality content in OSNs

In Sections 2.4 and 2.5, we studied the impact of influence and the role of semantics in the OSN analysis. Their combination can be used for assessing information dynamics as well as for the qualitative assessment of viral user-generated content. Here, we highlight the key points and considerations towards the definition and semantification of quality user-generated content.

According to the authors in (Chorley et al., 2015), metrics regarding retweets are the best quantitative indicators that show a preference for reading a tweet over another. From the readers' perspective, a tweet being retweeted several times is more attractive than a tweet with a lot of mentions. The authors conclude that the relationship among users and authors is the best qualitative indicator which has the strongest effect on the retweeting and reading processes. Retweets as quality indicators are also considered for measuring the appreciation of other users on the generated posts (Räbiger and Spiliopoulou, 2015). This attribute is highly used to calculate user influence.

The study in (Boyd et al., 2010) suggests that retweeting can also be characterized as a conversational infrastructure. According to the authors, a conversation "exists" either during a retweet where some new information can be added to the initial message, or when a single tweet is retweeted multiple times. The latter is interpreted as an action to invite new users into the conversation.

The rebroadcasting behavior of users is also studied in (Zhang et al., 2017) by examining three aspects: the role of content, content-user fit, and social influence. The content-user fit considers the interaction between the message content and the user interests whereas social influence measures the impact on the users' re-posting behaviors.

The work in (Zhang et al., 2016) describes a service for proposing news articles to Twitter users. It is based not only on terms which are mined from the users' and their friends' timeline, but it also incorporates additional factors that affect the interest of a user on a tweet, the number of retweets, and the influence of its publisher.

Personalization issues for recommendation purposes are also examined in (Li et al., 2017). The authors incorporate influential features (e.g. users' topic-level influence, topic information) and their relations among OSN users (e.g. retweeting behavior) for improving recommendation results in thematic categories. Social influence and its propagation can also be used as quality indicators in recommendation systems. In the work described in (Yuan et al., 2015), influence is considered as an attribute that propagates among users in OSNs. The proposed framework calculates the influence

that social relationships have on the users' rating behaviors, and incorporates it into recommendation proposals.

A study that evaluates the quality of articles on Wikipedia by investigating their usage on Twitter is presented in (Zangerle et al., 2015). This is achieved by analyzing three aspects, namely the language used in tweets in their referenced Wikipedia articles, the Twitted-related content features of such articles (e.g. URLs, hashtags, names of entities), and the correlation between the number of tweets/retweets and edits in their related articles. The authors discovered that the language of the tweets and the referenced Wikipedia articles are not always the same, mainly because of the low quality or the absence of equivalent entries in the user's native language. Moreover, it was found that the impact of a tweet/retweet about a certain topic is not related to crowd-sourcing-based metrics (e.g. edits, discussions) on the same Wikipedia topic.

In Section 2.4, we also described a framework (Dennett et al., 2016) that exploits influence for evaluating and enhancing communication issues between governmental agencies and citizens (OSNs users). The authors evaluate the quality of the agencies' responses with respect to the citizens' requests, analyze their sentiment and suggest influential users for agencies in order to obtain a new audience.

The authors in (Kumar et al., 2016) analyze and compare a variety of measurements in OSNs that affect the user influence. These were grouped under various criteria, namely neighborhood (i.e. number of influencers, personal network exposure), structural diversity (i.e. active community metrics), locality, temporal measures (i.e. retweet time delay), cascade measures (i.e. size, path length), and metadata (i.e. presence of links, mentions, hashtags). Moreover, based on several learning algorithms the authors propose methods to calculate the users' retweeting probability.

Another interesting methodology for measuring user influence based on the content quality is proposed in (Yu et al., 2016). Initially, users who disseminate quality content are considered those with high Follower-to-Followee ratio. According to the classification methodology, in the case of spam detection the users' influence is reduced. The authors introduced time as an important factor that affects the content influence and its probability of being viewed, retweeted or commented in different time zones.

Influential users are discovered in (Weng et al., 2010) by applying an algorithm which is an extension of PageRank. The algorithm takes into consideration both the topical similarity between users and their link structure. It is claimed that due to homophily (i.e. the tendency of individuals to associate and bond with others having similar interests) most of the "Follower-Following" relations are created.

Another approach that defines influence according to the behavior of directly related users (e.g. friends, followers, and so on) is presented in (Goyal et al., 2010). The authors proposed an "influenceability" score aimed at representing a user's susceptibility to be influenced by others. This score is built on the hypothesis that very active users perform actions without getting influenced by anyone. The study concluded that such kind of users should be considered as influencers in a network. Following the same rationale, in (Anagnostopoulos et al., 2008), the authors investigated social influence as a way of dictating users' behaviors in order to impose similar behaviors.

The authors in (Erlandsson et al., 2016) identify influential users by using association rule learning. As a machine-learning technique the specific approach investigates how

one item affects another by analyzing the frequency and the simultaneous appearance of certain items in a specific dataset. The authors assessed user participation in a post based on their previous interactions with other users on common posts. Toward this end, posts from Facebook pages were analyzed by extracting user actions, such as comments and likes. The authors claim that this technique allows the prediction of the participation of a particular user on a post discussion based on other users' activities.

In (Piškorec et al., 2016), the authors investigate whether peer and external influence can be inferred by using the user's friendship network. The experiment took place during an on-line voting procedure in Facebook. The analysis of the users' demographics and their votes showed a strong homophily among the communities and the friends' votes. The vote users influenced their friends to participate in this voting as well.

In (Yi et al., 2016), the authors analyze the so-called persuasion-driven social influence based on topic. Several influence measurements in terms of influence propagation for quantifying user-to-user influence probability incorporate the users' social persuasiveness. Based on the proposed metrics, the framework exploits the topical information, the user's authority and the characteristics of relationships (such as direct or indirect connections among users) among individuals.

In (Azaza et al., 2016), the authors propose an influence assessment approach for OSNs, by addressing limitations such as the lack of combined relationships and the uncertainty ignorance of existing ones. An influence graph is created to enable the observation of different relations and interactions, including retweets, mentions, and replies. Based on the belief functions theory, a general influence measure for a given user is established through an information fusion of the different relations. The proposed influence measure takes into account various interaction patterns in the graph, and considers derives influence from indirect nodes.

Finally, an important aspect in modeling the content quality in OSNs is the credibility of users since it is also strongly related to the trustworthiness of the information itself. One proposed solution is the crowd-based neutralized evaluation based on the accuracy, clarity and timeliness of information since its creation. When user-generated content is created, the topic/domain labeling enhances its credibility. In addition to personalization provision, each OSN user should also be able to create personalized filtering rules in order to select and evaluate the labeled content. By employing this mechanism, the community will eventually evaluate the topic/domain label credibility through collective intelligence and crowd-sourcing processes. Another solution includes personalization provision and topic/domain labeling tailored according to the users' information needs. The authors in (Haralabopoulos et al., 2016) propose several solutions which when combined with classic artificial intelligence and real-time data mining methods lead to a new form of social networking, which can serve as a qualitative and credible medium of information exchange.

2.7. Conclusions

In this Chapter, we reviewed two major aspects of OSNs, namely the online social influence and the role of semantics in OSNs (Sections 2.4 and 2.5 respectively), while in Section 2.6 we discussed how we may combine both aspects towards the qualitative

assessment and modeling of user-generated content. To perform a more detailed analysis and to adequately cover all the perspectives of the aforementioned aspects, we analyzed the reviewed works according to a proposed hierarchical classification scheme (Figure 2.1).

All of the related studies regarding influence measurements in OSNs conclude that the number of followers/friends a user has does not necessarily guarantee a high influence, despite affecting it to a certain extent. The most important factors that affect a user's influence can be categorized as:

- *User-oriented*: interaction with other users and similar activities (e.g. creating new messages), relationship details (number of followers, following users, friends), as well as structural network characteristics and attributes (e.g. position, shortest paths, closeness, eccentricity, centrality, and degree).
- *Content-oriented*: viral content (e.g. hashtags, mentions), identified user interests.
- *Quality-oriented*: where quality is measured by the user's social acknowledge and the degree of engagement with other users (e.g. the number of retweets/shares, favorites/likes, replies).

Usually, the users are highly influential mainly on some specific topics and less on others. We found that two types of influence exist; a topic-specific one, and a global one spreading through the entire network. Several recent studies propose that social influence should be incorporated into recommendation systems to leverage past behavior and latent relationships among users, as well as to improve their performance. In parallel, social semantics have been exploited in the analysis of users' behavior, interests and preferences, so as to help recommenders to suggest informative content, similar users, and other personalized information and others.

The literature that we have reviewed in this work has confirmed that influence and information flow are two interdependent concepts of OSNs, since they affect one another positively or negatively. Studies on dissemination of information have shown that the largest cascades tend to be generated by influential users who have many followers. Usually a large number of those -not so highly influential- followers initiate short diffusion chains which quickly merge into a large single structure. The dynamics of that information flow can be quantified by considering the following social diversities:

- User activity or passivity.
- User influence and susceptibility.
- User relationships in terms of interaction (e.g. mentions, replies) and friendship features.
- Reposting characteristics (e.g. volume, speed, time interval, number of hops).
- Homophily and entity similarity.
- Network attributes, structure and user topology.
- Content and structure of messages (e.g. topics, presence of URLs or hashtags, formality of language).

In addition, the study of social information spread (diffusion and propagation) is intrinsically connected to the problem of analyzing the modular structure of networks, known as community detection. The communities in OSNs promote certain topics and can be treated as the logical grouping of social actors that share common interests, ideas, or beliefs. There are two possible sources of information which can be used towards their detection: the network structure and the features and attributes of the nodes-users.

OSNs users often create messages that are characterized by the highly unstructured and informal language with many typographical errors, lack of structure, limited length, and high contextualization. Consequently, microblogging retrieval systems suffer from the problems of data sparseness and the semantic gap. To overcome those limitations and to contextualize the semantic meaning of microblog content, many recent studies focus on exploiting the use of social semantics and of user-generated content by identifying entities in them. These entities are used as keywords to indicate the topics of the messages, to describe real-life events, as well as to reveal behavioral patterns and building interest profiles, thus enabling the interrelation of semantically related terms and the social proximity or similarity between profiles and interests. Often, these entities are linked to knowledge bases (e.g. Google Knowledge Graph, DBpedia) or they are represented as concepts extracted from ontologies using Semantic Web vocabularies in order to transform unstructured data into Linked Data.

Chapter 3. Influence Properties and Metrics

3.1. Introduction

Microblogging is a form of Online Social Network (OSN) which attracts millions of users on daily basis, sharing hundreds of millions messages. Twitter is one of these microblog services. Their users vary from citizens to political persons and from news agencies to huge multinational companies. Obviously, some users are more influential than others. In that "ocean" of information, a challenging task is the discovery of the important actors who are able to influence others and produce messages of high social quality, importance and recognition. Therefore, a service is required for quantifying and measuring the value of that influence. To this end, we have created "*InfluenceTracker*[8]", a website where anyone can rate and compare the recent activity of any Twitter account. Specifically, we propose a novel measurement, namely "*Influence Metric*", the value of which derives from a social function incorporating i) the activity of a Twitter account (e.g. tweets, re-tweets, and replies), ii) its social degree (e.g. followers, and following) and iii) its qualitative content, reflected by other users' acknowledgement (e.g. retweets) and preferences (e.g. favorites).

Independent of the type of the user and of the degree of influence on others, all share the same need; their messages to reach as many users as possible. The messages, which are regarded as pieces of information, can be spread in two ways, either directly or indirectly. A case of direct message is when a company reveals information about a new product to its followers. When such a follower decides to share it among his/her own followers, i.e. to retweet, then that is a case of indirect information dissemination. Consequently, the tweets are viewed by accounts that are not being directly followed, resulting in the diffusion of information to users not targeted (to the followers of their followers). The same process can be continuously repeated.

To summarize, this chapter provides the following contributions:

1. We present "*Influence Metric*", a framework for calculating the importance and influence of Twitter accounts based on their activity, social degree, and quality.
2. We create "*InfluenceTracker*[8]", a publicly available website where anyone can rate and compare the recent activity of any Twitter account.
3. We propose a methodology, which describes the maximization of diffusion of information in OSNs.
4. We perform an experimental evaluation on real world data and validate that the selected social properties and characteristics are the appropriate for such a measurement to be based on.

The rest of this chapter is organized as follows. In sections 3.2 and 3.3 we discuss related work on the fields of measuring influence and information flow in OSNs. In section 3.4 we describe the proposed framework for measuring the importance and

[8] http://www.influencetracker.com

impact of Twitter accounts, while in section 3.5 we show how the dissemination of information in Twittersphere can be calculated. Finally, in section 3.6 we present experimental evaluation on real case scenarios, along with their results and assessment.

3.2. Measuring Influence in OSNs

The topic of measuring user influence on social networks, as well as the identification of opinion leaders on them is not new. It spans over a wide area of sciences, covering from social sciences to viral marketing and from daily communication to OSN platforms. In the bibliography there is no clear definition of the "influential user". Hence, the term "influence" has multiple interpretations. Consequently, emerging influence metrics are continuously varying with each of them using different criteria. Despite this variation, all the related studies do share a common result, which is that the most active or popular (that with many followers) users are not necessarily the most influential ones.

The study in (Anger and Kittl, 2011) proposes the "Social Networking Potential" as a quantitative measurement for discovering influential users in Twitter, and suggests that having a large number of followers does not guaranty high influence. The methodology is based on the number of tweets, replies, retweets, and mentions of an account.

Influence in terms of activity or passivity for Twitter users is studied in (Romero et al., 2011b). To conduct this study, a corpus of tweets is utilized including at least one URL, their creators and their followers. The derived measurement is based on the "Follower-Following" relationships of the users, in addition to retweeting patterns. As most studies in this area, it is stated that the number of followers a user has is a relatively weak predictor of the maximum number of views a URL can achieve. As our work has shown (Razis and Anagnostopoulos, 2014a), the number of followers an account has does not guarantee the maximum diffusion of information in Twitter. This is because, in order to achieve high levels of diffusion, your followers should not only be active, but they should also have a high probability of retweeting, thus transmitting the messages they receive to their followers.

The authors in (Cha et al., 2010) introduce three types of influence for Twitter users, namely "In-degree" (number of followers), "Retweet" (number of user generated tweets that got retweeted) and "Mention" influence (number of times the user is mentioned in other users' tweets). For calculating these influence types, the users should post at least ten messages. According to the authors, the "Retweet" and "Mention" influence correlate well with each other, while the "In-degree" does not. Consequently, the study concludes that the most followed users are not necessarily influential.

Influential users are discovered in (Weng et al., 2010) by applying an algorithm as an extension of PageRank, which takes into consideration both the topical similarity among users and their link structure. It is claimed that due to homophily, that is the tendency of individuals to associate and bond with others having similar interests, most of the "Follower-Following" relations appear. This work also suggests that the active users are not necessarily influential.

Another approach that defines influence according to the behavior of the directly related users (e.g. friends, followers) is presented in (Goyal et al., 2010). The authors propose an "influenceability" score aimed at representing a user's susceptibility to be influenced by others. This score is built on the hypothesis that very active users perform actions without getting influenced by anyone. The study concludes that such kind of users should be considered as influencers in a network

The study in (Boyd et al., 2010) suggests that retweeting can also be characterized as a conversational frame. According to the authors, a conversation "exists" either during a retweet where some new information can be added to the initial message, or when a single tweet is retweeted multiple times. The latter is interpreted as an action to invite new users into the conversation. In our work (Razis and Anagnostopoulos, 2014b) we propose that other users' actions on the messages, such as retweets and marking as favorites, can be regarded as qualitative indicators over their content.

In (King et al., 2013) the "*t*-index" metric is proposed, aiming at measuring the influence of a user on a particular topic. This metric is based on the *h*-index factor and indicates the number of times a user's tweet on a certain topic has been retweeted. The authors suggest that a high influence on one topic does not necessarily mean the same on other topics. In our work (Razis and Anagnostopoulos, 2014a) we also propose that the incorporation h-index metrics, namely "Retweet h-index" and "Favorite h-index", over the proposed Influence Metric can improve its accuracy, by reflecting the impact of OSNs users (Razis and Anagnostopoulos, 2014b).

A graph-based approach for the identification of influential users in OSNs is presented in (Beiming and Ng, 2012). A created graph represents the relationships among the tweets and the users. The more implicit or explicit relationships among tweets exist for a user, the more influential the user is. This work, which is one of the first in the area, considers only the number of tweets as an influential factor, which has been later proved to be a weak predictor.

A framework exploiting influence for evaluating and enhancing communication issues between governmental agencies and citizens in OSNs is proposed in (Dennett et al., 2016). The aim here is to evaluate the quality of the agencies' responses with respect to the citizens' requests, to analyze the citizens' sentimental attitude and their subsequent behaviors, and to suggest influential users to the agencies in order for the agencies to obtain new audience. To achieve these goals, several components are incorporated into the framework. These components detect the demographics of the followers, their locations, topics of interest, and sentiment.

Finally, the authors in (Hassan et al., 2016) propose a different kind of influence called the "susceptibility to influence". This metric estimates how easily a Twitter user can get influenced. The proposed metric utilizes the user's social interactions that depend on three factors namely, activity, sociability and retweeting habit. Activity reflects the user's tendency to interact with friends and consequently the chance to become influenced by them, while sociability corresponds to the users' social degree among their activities, implying that interactions with more friends results in a wider diversity of topics and interests.

All the related studies have shown that the most active users or the ones with the higher number of followers are not necessarily the most influential. As described in Section 3.4, our Influence Metric depends on a set of factors, where the account activity is only one of them. Simply put, as the authors in (Srinivasan et al., 2014)

state, enormous influence may spring from lesser known persons, while the "celebrities" may not be influencers.

Contrary to the aforementioned studies, for the calculation of our Influence Metric we neither set a lower threshold on the number of the user-generated tweets, nor we utilize only a specific subset of tweets that fulfill certain criteria (e.g. those containing URL). Our proposed metric is concentrated on the characteristics of the Twitter accounts; consequently all of them can be used as seed for the calculation of Influence Metric, thus differentiating our work in respect to the related literature.

A detailed analysis of studies describing influence measurements which are based on other aspects, such as hyperlink-based algorithms (e.g. PageRank or related ones) and machine-learning techniques can be found in Section 2.4.1.

3.3. Information Flow and Influence in OSNs

Information flow is vital in all kinds of networks (e.g. social, digital, or computer), and this flow can be affected by the actions or properties of their actors and the sets of dyadic relationships between them. Influential users determine the virality of information and, specifically, how such information is spread.

In several works, an individual's influence is calculated based on the volume of information spread over a network. For each influenced node an algorithm calculates how many others nodes are affected. The algorithm is based on the hypothesis that the number of newly influenced nodes is determined by the previously influenced ones. The study described in (Yang and Leskovec, 2010) suggests that the diffusion of information is governed by the influence of individual nodes. Similarly to (Barbieri et al., 2013), the proposed models are considered as stochastic processes in which, according to probabilistic rules, information is spread from a node to its neighbors. In a similar way, the study described in (Kimura and Saito, 2006) aims at the identification of influential nodes by calculating the expected number of influenced ones.

The authors in (Huang et al., 2013) propose an extension of PageRank for measuring influence. They apply their extended PageRank approach on a graph of retweets and user relationships and consider social diversity of users and transmission probabilities of the messages based on the hypothesis that users inherit influence from their followers. The aim is to explore whether individual characteristics and social actions as well as influence propagation patterns are factors capable of influencing other users.

The "retweet" functionality and the retweet counter can be considered as a factor for measuring the "interestingness" of a user's tweets (Naveed et al., 2011a). Based on that, the resulting spread of information is examined in (Kwak et al., 2010). The authors propose that the retweet counters are indicators of the popularity of the messages and of their authors. According to the study, as soon as a post gets retweeted, it will be nearly immediately spread up to four hops away from the source, thus resulting in a rapid diffusion after the first retweet. Three measures of influence namely, the number of followers, PageRank, and the number of retweets, are further compared and evaluated. The results indicate that, in contrast to the third measurement, the first two provide similar rankings of influential users, indicating a gap in the influence derived from the number of followers and the popularity of the

tweets. In (Romero et al., 2011b), the derived influence metric correlates the information propagation with Twitter user's retweeting behavior, and it is used for measuring the activity or passivity of those users. The "retweet" function as a metric is also utilized in our proposed "Influence Metric" (Section 3.4).

Finally, there are other works ((Lerman and Ghosh, 2010), (Yang and Counts, 2010), and (Bakshy et al., 2012)) that investigate the discovery of information propagation flows in OSNs. Information diffusion in two social networks, namely Digg and Twitter, is studied in (Lerman and Ghosh, 2010). According to the study, information flow and spread are affected by the structure of these networks. Information in networks with sparse and poorly interrelated structure (e.g. Twitter) reaches nodes slower in comparison to networks with a dense structure (e.g. Digg), in which information spreads faster. Due to the structure of Twitter, information may spread at a lower pace, but it maintains its diffusion at the same rate as time passes, thus penetrating the network further.

In (Yang and Counts, 2010), a dissemination network is constructed based on user mentions, with constraints on topical similarities in the tweets. The authors suggest that the frequency of mentioning a user is among the strongest predictors of information spread.

Similarly, the authors in (Bakshy et al., 2012) examine information propagation on Facebook that provides insights into information shared by friends. They found that the individuals who are aware of these insights are more willing to re-share the information faster, compared to those who are not. Although these strong relationships are more influential separately, the weaker bonds, exceeding those in numbers, are responsible for the propagation of information.

Many of the presented studies on the information flow in OSNs aim at identifying nodes of high influence as responsible for affecting neighboring ones to behave the same way, in terms of spreading information of similar content. As the results show in Section 3.5, our proposed Influence Metric (defined in Section 3.4) succeeds in indentifying the nodes that result in the maximization of information diffusion in an OSN.

As already mentioned in Section 2.2, in the literature the terms "propagation" and "diffusion" are often used interchangeably when referring to information flow and dissemination. In this thesis, we explicitly examine them separately, as diffusion defines the spread of information from a starting node towards the rest of the network, while propagation takes into consideration the intermediate nodes as well, which receive, process, and further decides how to handle information.

3.4. Influence Metric: Our proposal

Twitter accounts form a social network. If depicted in a graph, they are represented by nodes. Edges that connect these nodes are the relations of "Follower-Followee" instances. Even if some accounts are more influential than others, the influence measurement should not merely depend on the number of "Followers", even if this number is big enough. In case that the number of "Followees" is larger, then the user could be characterized as a "passive" one. These type of users are regarded as the one who are keener on viewing or being informed through tweets rather than composing new ones. Therefore, a more suitable factor is the ratio of "Followers to Followees"

(*FtF* ratio). However, this ratio is also not sufficient. Another important factor is the tweet creation rate (*TCR*). For example, let us see the case where two accounts have nearly the same *FtF* ratio. Obviously, the account with the higher *TCR* tends to have a higher impact in the network. In our methodology, and in order to calculate that rate, we process the latest 100 tweets as provided from the Twitter API. That helps us to keep the values of *TCR* (and consequently the *Influence Metric*) dynamic, as this value depends on the most recent activity of the accounts in Twitter. In order to maximize the precision of the metric, the timeframe of its calculation is measured in hours, instead of days.

Each tweet is associated with several other kinds of the information presented in *InfluenceTracker.com*. Two of them are the "Retweets" and "Favorites" counts, which represent how many times a Tweet has been retweeted as well as how many times it is marked as favorite by other users respectively. In our methodology, we utilize these counts in order to calculate the *h*-index of the "Retweets" and "Favorites", over the last 100 tweets of an examined account. The aim of these measurements is to provide a quality overview of the tweets of a Twitter account in terms of likeability and impact in Twittersphere. These indexes are based on the established *h*-index (Hirsch, 2005) metric and are called *"Retweet h-index - Last 100 Tweets"* and *"Favorite h-index - Last 100 Tweets"*. The most important factor regarding them is that they reflect other users' assessment of the content of the tweets.

Consequently, a Twitter account has a *"Retweet h-index - Last 100 Tweets"* equal to *h*, if *h* over the last *Nt* tweets have at least *h* retweets each, and the remaining *(Nt - h)* of these tweets have no more than *h* retweets each (max. *Nt*=100). This can be interpreted as follows: at least *h* tweets have been retweeted at least *h* times. Thus, we consider that this retweeting action results in the generation of at least *h*h* new tweets, which have to be attributed to the account that initially posted them.

Prior to incorporating this amount of new tweets into the equation of the Influence Metric, we employ a calculation mechanism for avoiding outliers. Moreover, we introduce a value called "Adjusted Tweets" which is defined in Equation 3.1.

$$\begin{gathered}\text{Adjusted Tweets} = \mathrm{a} \times 10^{\mathrm{b}}, \\ \text{where } 0 < a < 100, 0 \leq b \leq 3 \\ \text{and } \mathrm{a} \in \mathcal{R}, \mathrm{b} \in \mathbb{Z}.\end{gathered} \qquad (3.1)$$

The "Adjusted Tweets" are actually a form of expressing the *h*h* value. Characteristic examples of this metric are provided in Table 3.1. The factor "a" is transformed into a number lower than 100 (third column in Table 3.1), which is the maximum number of tweets considered for each examined account. Then, this number is divided by 10 (green numbers of fourth column in Table 3.1). The resulting quotient is combined with the Order of Magnitude (OOM) of the *h*h* (red numbers of fourth column in Table 3.1), which is represented by "b", thus forming the "Adjusted Tweets Number" according to Equation 3.1.

Table 3.1: Examples presenting the calculation of the "Adjusted Tweets" value

RT *h*-index	*h***h*	Transformed as	Calculation Process	Adjusted Tweets
0.3	-	0.3 * 10^0	0.3/10, 10^0	0.03
2	4	4 * 10^0	4/10, 10^0	0.4
6	36	36 * 10^0	36/10, 10^0	03.6
15	225	22.5 * 10^1	22/10, 10^1	12.2
45	2,025	20.25 * 10^2	20/10, 10^2	22
80	6,400	64 * 10^2	64/10, 10^2	26.4
100	10,000	10 * 10^3	10/10, 10^3	31

As already mentioned, the tweets generated from the retweeting process have to be attributed to the account that initially posted them. To achieve this, the value of the "Adjusted Tweets" is added to the 100 tweets retrieved from the account, as defined in Equation 3.2. The *FtF* ratio is placed inside a base-10 log for avoiding outlier values. Moreover, this ratio is added by 1, so as to avoid the metric being equal to 0 in case where the values "Followers" and "Followees" are equal. For example, if an account has 10,000 followers then OOM equals to 4.

$$\text{Influence Metric} = \frac{\text{tweets}_{\text{k}} + \text{AdjustedTweets}_{\text{k}}}{\text{Hours}_{\text{since k}_{\text{th}}\text{tweet}}} * \\ * \text{OOM}(\text{Followers}) * \log_{10}\left(\frac{\text{Followers}}{\text{Followees}} + 1\right), \tag{3.2}$$

3.5. Measuring Information Spread

An important functionality provided by Twitter is the "retweet", allowing users to repost a received tweet to their Followers. Consequently, the tweets are viewed by accounts that are not being directly followed, resulting in the diffusion of information to users not targeted (to the followers of their followers). The same process can be repeatedly take place by the new viewers of the message and so on.

The most important factor which affects the transmission of the tweets is the followers' probability of retweeting. The higher this value, the higher the probability of transmitting tweets to other users, initially not targeted by the source. Another factor affecting the transmission of the tweets is the followers' TCR. The value of this rate includes both the accounts' generated tweets, as well as their retweets. The final dependency of that measurement is the "TCR of Follower to TCR of Account" ratio. Increased values of this ratio lead in bigger flows of tweets between these Twitter accounts. The Tweet Transmission (TT) metric depends on all of the aforementioned

characteristics of the directly related accounts and it is defined in Equation 3.3. A user's retweeting (RT) probability is based on the latest 100 retrieved tweets.

$$\text{Tweet Transmission} = \frac{\text{TCR}_{n+1}}{\text{TCR}_n} * \text{RT}_{n+1}, \text{where } n \geq 0, n \in Z, 0 \leq \text{RT} \leq 1 \quad (3.3)$$

3.6. Evaluation

In this section, we present and analyze the results and the evaluation concerning the calculation of the importance and influence of a user in an OSN, the effect the selected influential properties have, and the framework regarding the maximization of diffusion of information. As a case study, we evaluate six real Twitter accounts. Three belong to political persons (@AdonisGeorgiadi, @IliasKasidiaris, and @PanosKammenos), one belongs to the Hellenic Fire Brigade (@Pyrosvestiki), and the rest belong to a Greek news media channel (@SkaiGr) and to the international information network of activists and hacktivists called Anonymous (@YourAnonNews). The experiments took place between 14/12/2013 and 31/1/2014. For each account four separate samplings were made, during which the number of the followers and the top-k accounts were gradually increased.

Moreover, in an effort to directly evaluate and compare our metric with other approaches we used a well-known service, called Followerwonk[9], which provides the "Social Authority" value, another measure of how influential an account is. To this end, we randomly selected nearly 11,500 Twitter accounts and then we compared the ranking positions as well as the average ranking differences in all the positions levels for both services (InfluenceTracker and Followerwonk) (Section 3.6.1.2).

3.6.1. Evaluation of Influence Metric

3.6.1.1. Evaluation of Influential Properties

In this section, we present the Influence Metric measurements with respect to the examined Twitter accounts. In Table 3.2 we also provide the sampling date, as well as other metrics.

As we can see, the Influence Metric measurement directly depends on the accounts' activity, which is measured by the TCR value. The account's "@SkaiGr" influence value during the first three samplings (SG1 to SG3) is approximately the same (nearly 35 Millions). However, during the fourth sampling (SG4) that value was almost the half. This was caused by the fact that the TCR value dropped to half, despite that the "Followers to Following" ratio slightly increased. In the case of the account called "@YourAnonNews", during the first 3 samplings (YAN1 to YAN3) the Influence Metric value approximately equal to 341 Million. During the last sampling, YAN4, this value dropped to approximately 329 Million. This derives from the smaller value of "Followers to Following" ratio (the amount of following accounts increased during the period of the last sampling).

We should note here, that for the calculation of the Influence Metric, we consider the latest 100 of the accounts' tweets directly from the Twitter API. This enables the

[9] http://followerwonk.com/

measurement to be dynamic and in accordance to the latest trend activity of the examined Twitter account.

Table 3.2: The Influence Metric measurement and the Twitter related characteristics of the examined Twitter accounts

ID	Username	Date	Influence	TCR	Followers	Following
AG1	@AdonisGeorgiadi	14/12/2013	126,857.416	15.50	33,410	3,574
AG2	@AdonisGeorgiadi	18/12/2013	112,929.569	11.11	33,566	3,576
AG3	@AdonisGeorgiadi	29/12/2013	511,537.359	50.00	34,164	3,579
AG4	@AdonisGeorgiadi	16/01/2014	148,166.219	14.29	35,430	3,584
IK1	@IliasKasidiaris	16/12/2013	26,686.871	1.11	14,148	56
IK2	@IliasKasidiaris	19/12/2013	26,927.978	1.12	14,150	56
IK3	@IliasKasidiaris	26/12/2013	25,492.531	1.06	14,172	56
IK4	@IliasKasidiaris	26/01/2014	23,840.975	0.99	14,278	56
PK1	@PanosKammenos	17/12/2013	63,708.939	3.33	33,889	419
PK2	@PanosKammenos	21/12/2013	56,266.498	2.94	33,940	419
PK3	@PanosKammenos	29/12/2013	46,724.779	2.44	34,029	419
PK4	@PanosKammenos	12/01/2014	41,621.203	2.17	34,274	419
P1	@Pyrosvestiki	01/01/2014	23,516.011	0.62	18,619	3
P2	@Pyrosvestiki	30/01/2014	23,894.273	0.63	18,612	3
P3	@Pyrosvestiki	31/01/2014	23,516.156	0.62	18,620	3
P4	@Pyrosvestiki	31/01/2014	23,516.011	0.62	18,619	3
SG1	@SkaiGr	17/12/2013	35,356,300.107	100.00	178,446	52
SG2	@SkaiGr	21/12/2013	35,363,204.477	100.00	178,730	52
SG3	@SkaiGr	31/12/2013	35,380,441.726	100.00	179,441	52
SG4	@SkaiGr	01/01/2014	17,733,148.729	50.00	179,505	51
YAN1	@YourAnonNews	18/12/2013	341,594,730.673	100.00	1,185,201	455
YAN2	@YourAnonNews	23/12/2013	341,102,758.175	100.00	1,184,723	460
YAN3	@YourAnonNews	27/12/2013	340,808,348.148	100.00	1,184,390	463
YAN4	@YourAnonNews	24/01/2014	328,801,969.528	100.00	1,189,204	613

3.6.1.2. Evaluation against Followerwonk

In order to directly evaluate our metric, we randomly selected nearly 11,500 Twitter accounts and then we compared the ranking positions, as well as the average ranking differences in all the positions levels for both services (InfluenceTracker and Followerwonk). In Figure 3.1, the horizontal axis value of each point corresponds to the ranking position assigned by InfluenceTracker, while the vertical-axis value corresponds to the respective ranking assigned by Followerwonk. We observe that the average position ranking difference is 1,476 (the black linear trendline is defined by the function $y=0.783*x+1476$). Finally, in comparison to the ideal curve (red line that is defined by $y=x$) we noticed that outlier values are equally distributed between the higher and the lower ranking positions assigned by our service.

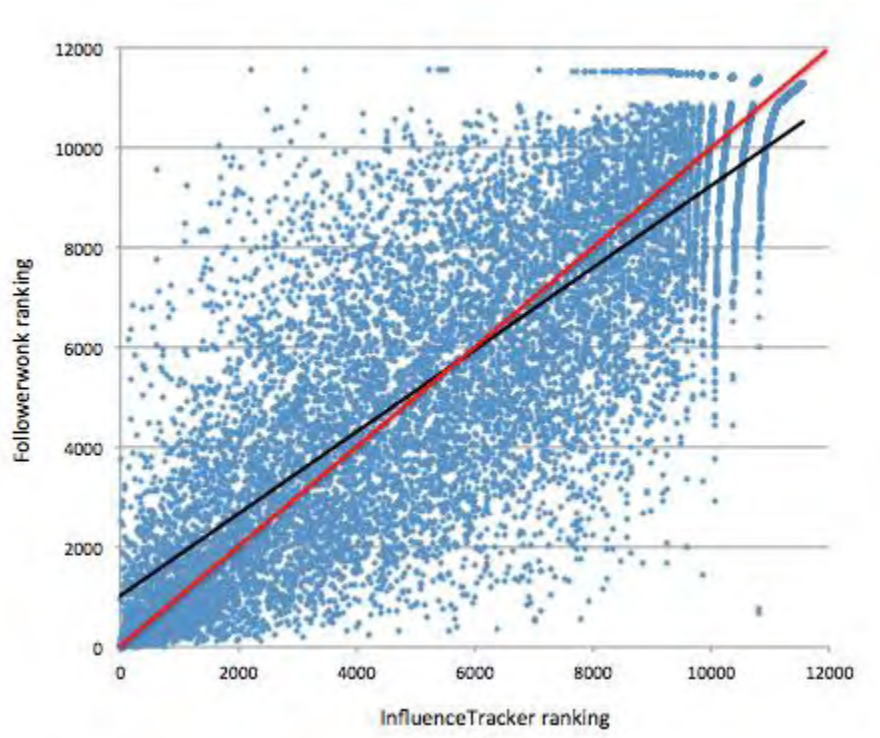

Figure 3.1: Evaluation of our metric in comparison to the Followerwonk service

In order to further compare and evaluate the quality of the Influence Metric, we extended the coverage of the collected data of the aforementioned accounts to other aspects of OSNs as well. Our aim is twofold: a) to evaluate the individual properties on which Influence Metric is based (Section 3.4), and b) to further compare our metric against Followerwonk. To this end, we collected or calculated the following 18 properties:

1. Influence Metric value,
2. Social Authority value,
3. number of followers,
4. number of following,
5. number of tweets,
6. number of tweets per day,
7. lifespan of account,
8. last-100 h-index retweet value,
9. daily h-index retweet value,
10. last-100 h-index favorite value,
11. daily h-index favorite value,
12. retweet percentage,
13. reply percentage,
14. number of mentions,

15. number of replies,
16. number of hashtags,
17. number of URLs, and
18. number of photos.

In order to facilitate the evaluation process, we classified all the accounts (approx. 11,500) into 28 categories. Thus, we have ones that represent the top 0.1% to 1% of the accounts having the highest Influence Metric value with an increasing step of 0.1% (10 categories), the top 2% to 10% with an increasing step of 1% (9 categories), and the remaining accounts with an increasing step of 10% (9 categories). For each category, we calculated the average values of the aforementioned 18 properties based on the corresponding accounts.

Figures 3.4 to 3.10 present the distribution of the values of the properties which are used for the calculation of Influence Metric (Section 3.4). The horizontal axis of each point corresponds (in logarithmic scale) to the number of accounts belonging to a category, while the vertical axis corresponds to the value of a property for that category. The black curved lines in the figures present the power trendline of each distribution.

Figures 3.4 to 3.6 present the distribution of the basic properties, such as the number of followers, following and daily tweets respectively. Figures 3.7 to 3.10 present the distribution of qualitative properties, namely *last-100* and *daily* h-index retweet, as well as the *last-100* and *daily* h-index favorite respectively. Finally, Figures 3.2 and 3.3 present the distribution of the accounts' importance, as measured by the InfluenceTracker and Followerwonk services.

As it can be seen from the trend-lines, the properties which are used for the calculation of Influence Metric (Figures 3.4 to 3.10) follow a power law distribution. The long tails of the distributions are clearly visible on the right hand side, while on the left hand side the fewer accounts having the greatest values for those properties are located. The distribution of Followers (Figure 3.4) can be regarded as the baseline of the power law distribution, as it almost overlaps with the trendline. The rest of the distributions approach the baseline, but mostly the Daily h-index Retweet and Favorite.

By considering the aforementioned, we come to the following conclusions:

a. We validated that our Influence Metric is based on the appropriate social properties, namely: i) the activity of a Twitter account (number of tweets per day), ii) its social degree (number of followers and following) and iii) its qualitative content, reflected by other users' acknowledgement (h-index retweet value) and preferences (h-index favorites value).
b. By comparing the distributions and the trendlines of Influence Metric (Figure 3.2) and Social Authority (Figure 3.3), it can be clearly seen that IM has a better adaptation to the power law distribution, namely "Few accounts have a very big score and too many very small". Moreover, the IM values spread nearly from the start of the y axis, in comparison to the Social Authority the lowest value of which lies at a very high position in that axis. Consequently, our metric seems to be much more accurate and well-defined.

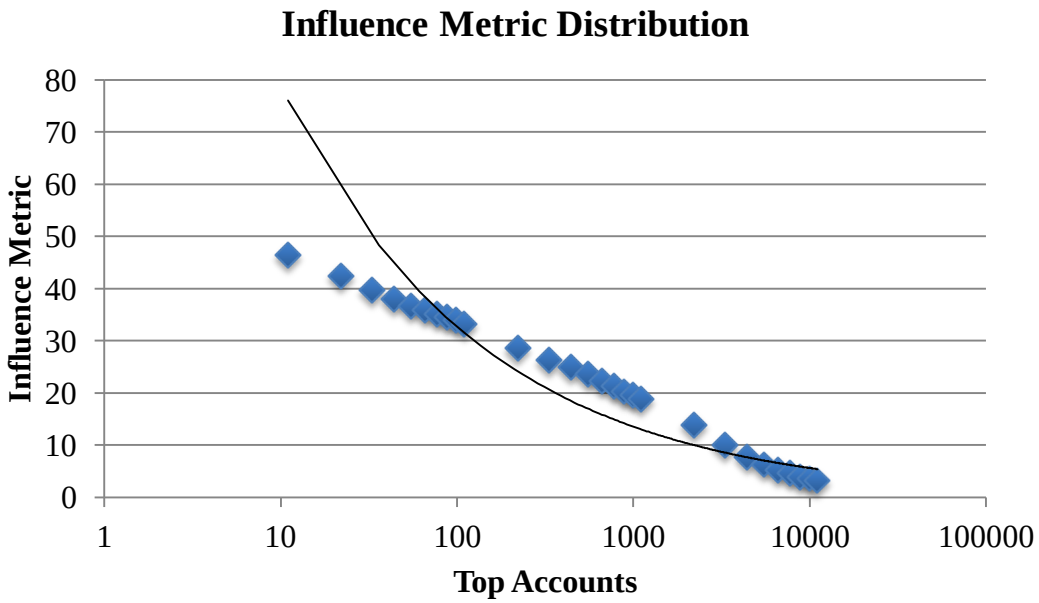

Figure 3.2: Distribution of the Influence Metric value

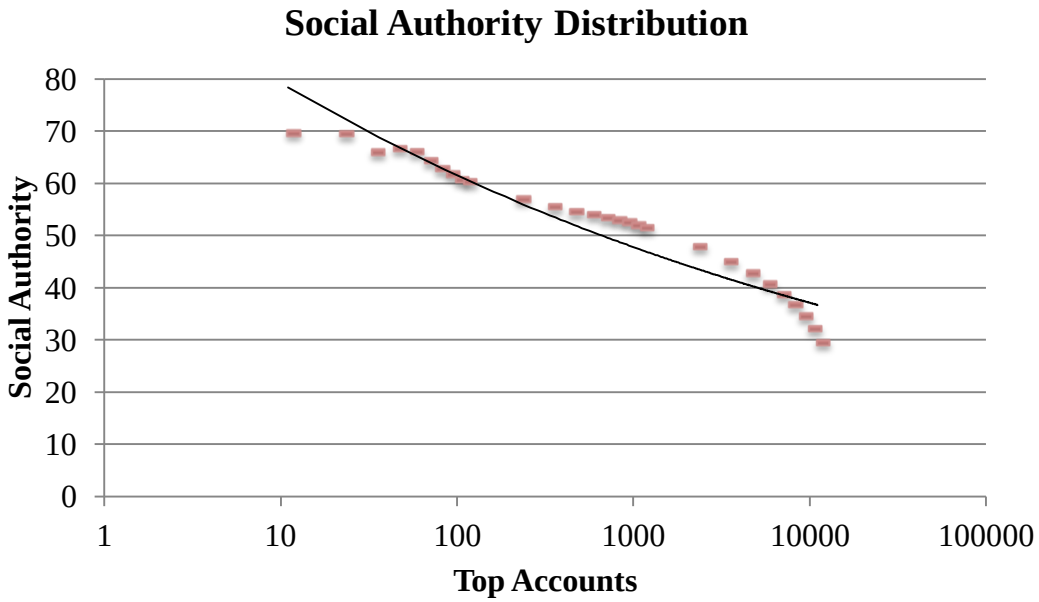

Figure 3.3: Distribution of the Social Authority value

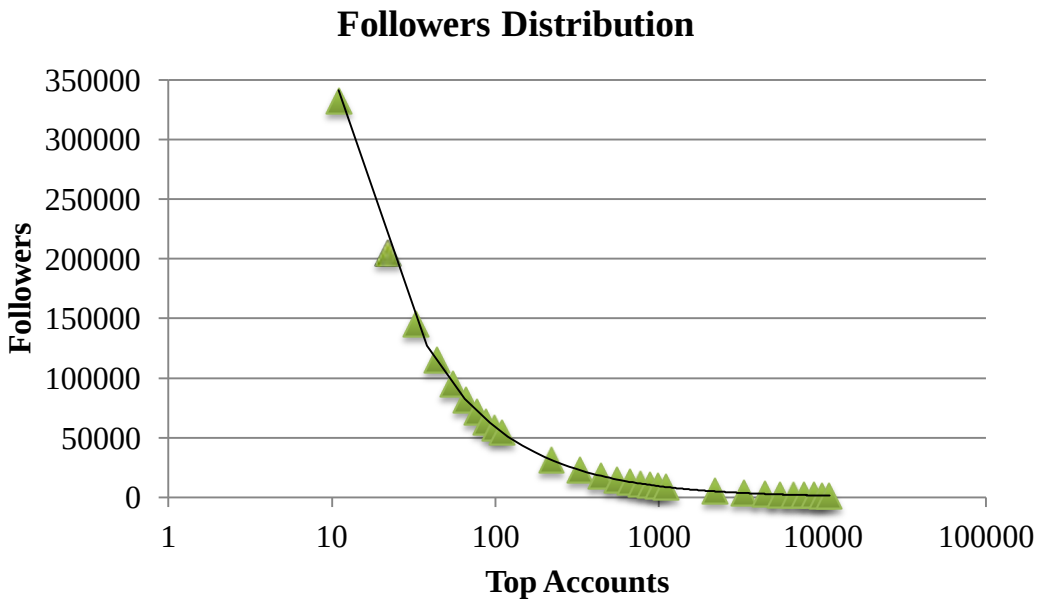

Figure 3.4: Distribution of the number of Followers

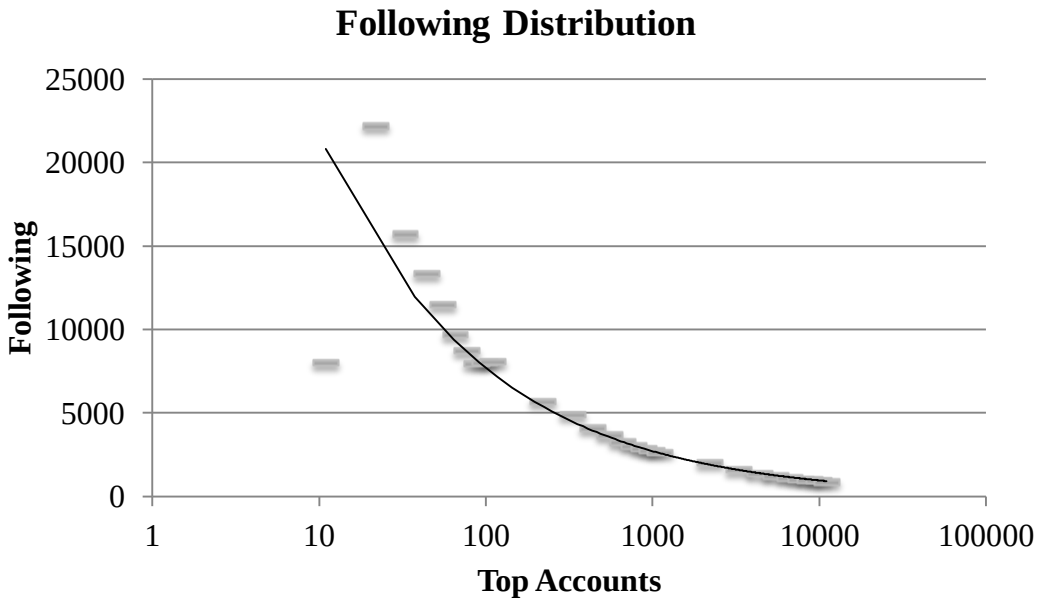

Figure 3.5: Distribution of number of the Following

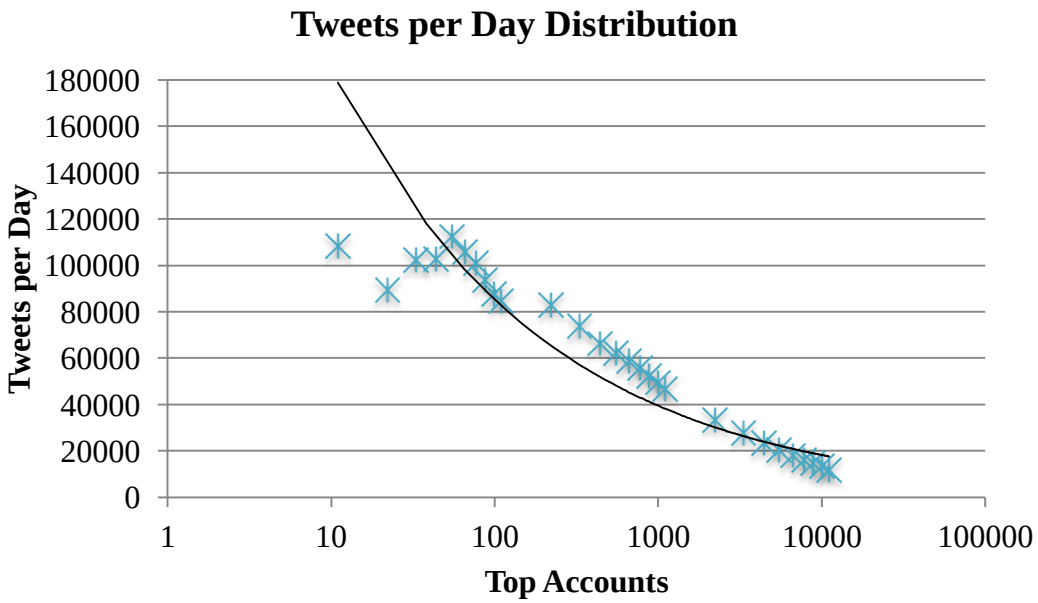

Figure 3.6: Distribution of number of the daily Tweets

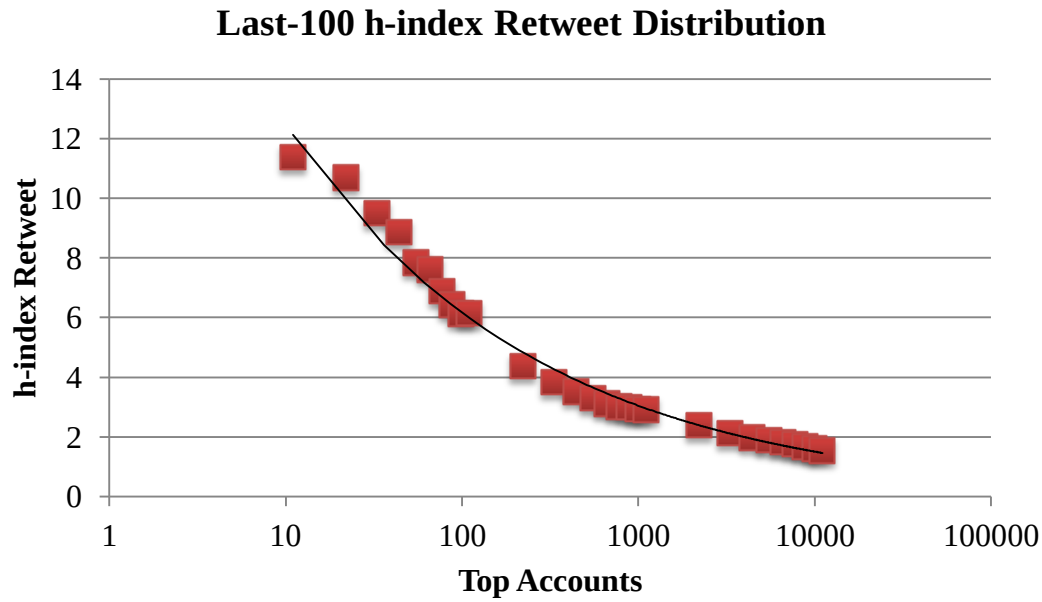

Figure 3.7: Distribution of the h-index Retweet value

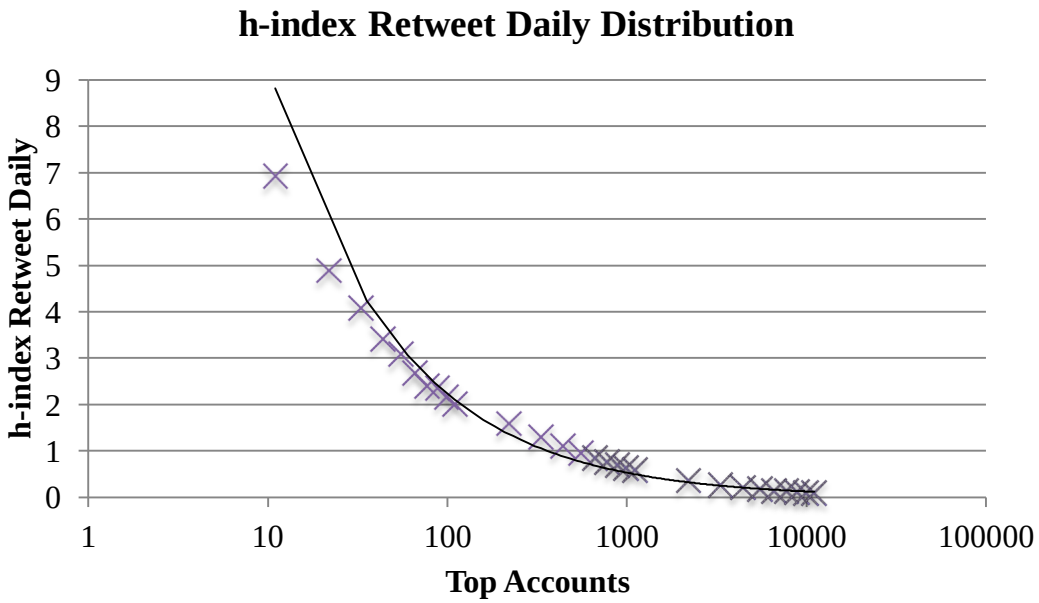

Figure 3.8: Distribution of the daily h-index Retweet value

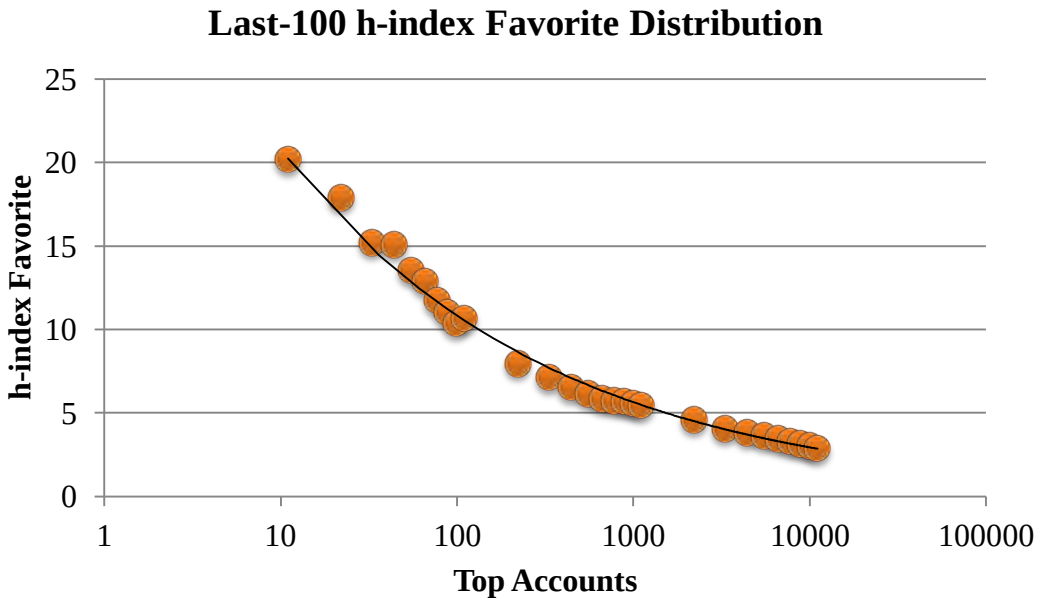

Figure 3.9: Distribution of the h-index Favorite value

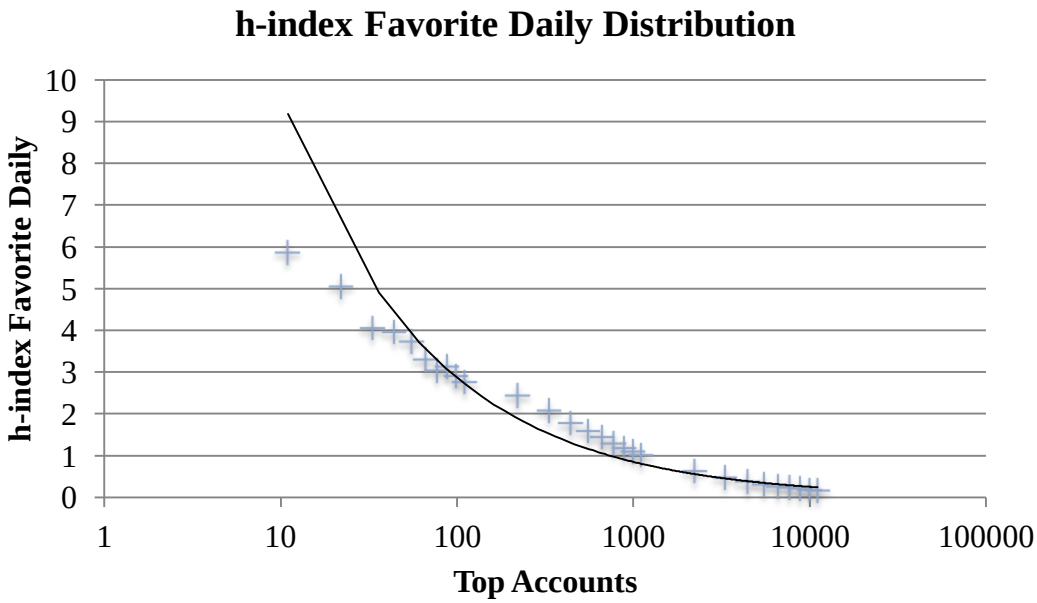

Figure 3.10: Distribution of the daily h-index Favorite value

3.6.2. Evaluation of Tweet Transmissions

3.6.2.1. Experimental Setup

In order to evaluate the diffusion-related metrics we employ the evaluation framework proposed at Figure 3.11.

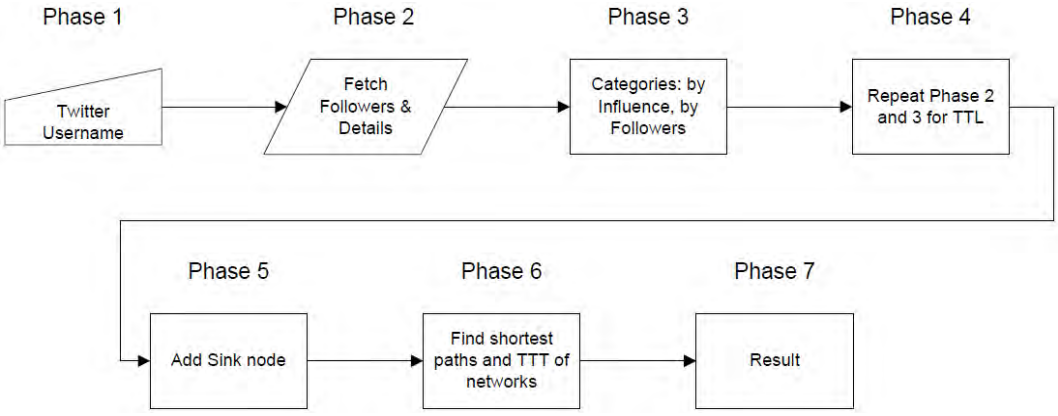

Figure 3.11: The seven phases of the proposed framework

The framework is split into seven phases, as presented in Figure 3.11. During Phase 1 of the process, the Twitter account under examination is selected. In Phase 2, we fetch a large number of followers (N_f) and their Twitter-related characteristics. These are necessary in order to calculate their Influence Metric measurement (Equation 3.3).

In Phase 3, the followers of the examined account are placed in two categories. The first one is classified by the value of Influence Metric, while the second one by the absolute number of followers each follower has. Both of these categories are sorted in

descending order. After that, we select the top-k follower accounts of these two categories.

For these top-k accounts, Phases 2 and 3 are repeated in the same way. This process is continued until a specified distance threshold (layers) between the Twitter accounts is reached (Phase 4). In computer networks, this distance is expressed by the Time-To-Live value (TTL) and corresponds to the amount of hops between different nodes a transmitted packet can circulate before being rejected by the network. For the purposes of this work this threshold is set equal to 3.

The examined account, all its followers, as well as their relations and characteristics are modeled as a separate network. Nodes depict accounts, while edges depict their relations containing specific attributes. As a result of this process, two network structures of the initial account and the followers of followers are created. One depicts the top-k accounts according to their Influence Metric value, while the other the top-k accounts according to the number of their followers. An example of such a graph is presented in Figure 3.12, where a 3-layered structure graph is displayed. The blue node represents the initial examined account. This account is connected to the yellow nodes, which stand for the top-3 followers measured either by the Influence Metric or by the number of their followers (1st layer of distance). The process is iteratively continued with these nodes. The green and red nodes, 2nd and 3rd layer respectively, represent the followers of the previously examined followers and so on. We should note here, that a node can be connected with others, independently of whether they belong to the same layer or not.

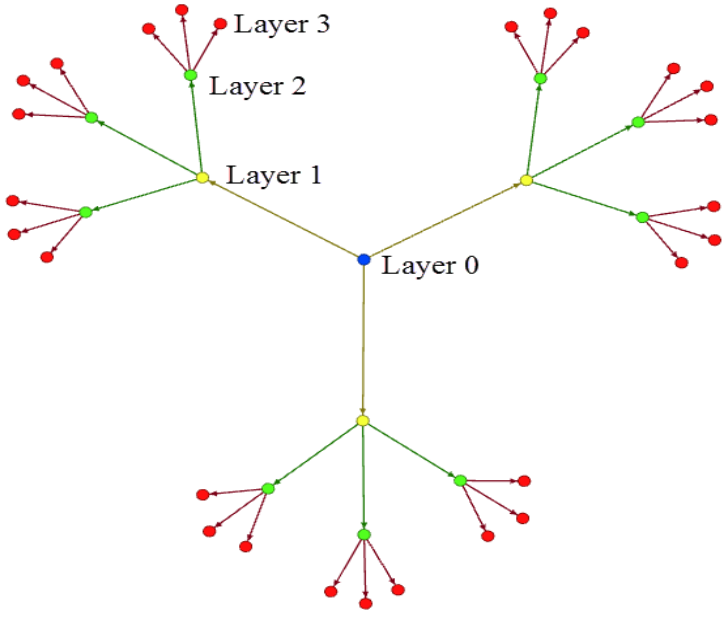

Figure 3.12: A 3-layered structure graph of the initial account and the top-3 followers of followers

During Phase 5 an ending node (sink) is added to each of the two generated networks. This node is connected with all the accounts-followers of the last layer. These are the red nodes of Figure 3.12 which belong to the 3rd layer. That results in a fixed starting and a fixed ending network point. Figure 3.13 presents the network illustrated in Figure 3.12 including the sink node (black node in the center).

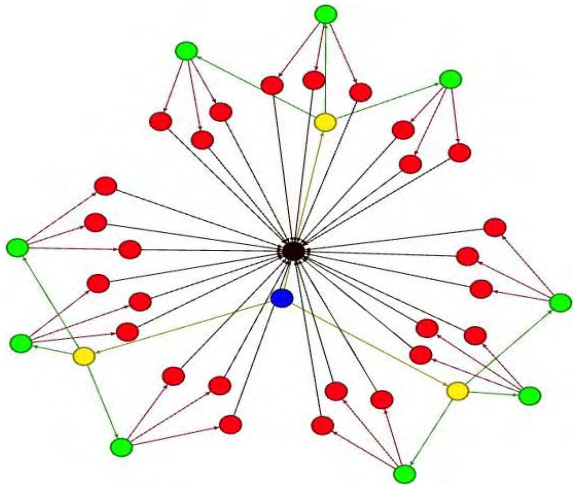

Figure 3.13: A 3-layered structure graph with a sink node

When all the previous phases are completed, the sixth and final Phase is initiated. Its purpose is to discover all the paths which contain exactly one node belonging to every layer, thus consisting of exactly 4 steps, starting from the initial examined account (blue node) and ending to the sink (black node). Furthermore, that number of steps ensures that any possible loops will be avoided during the traversal of the network from the initial account to the sink. A possible case of loop is when the examined Twitter account appears as a follower of another account. In such a case, the examined account could also appear at the first (as a yellow node) or the second layer (as a green node).

The Tweet Transmission (TT) value, presented in Equation 3.3, is calculated for each layer of every shortest path. Then, the TT value of the shortest path for all the layers is calculated. This process is repeated until the TT values of all the shortest paths of the two networks are computed. The network with the higher Total Tweet Transmission (TTT) value is considered the one with the higher disseminated information.

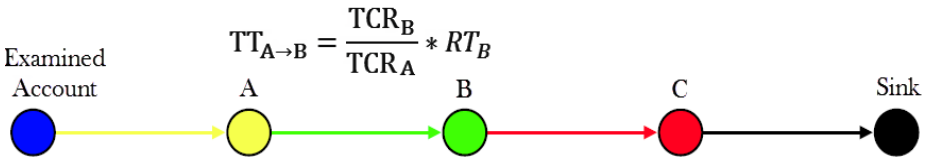

Figure 3.14: Calculation of a TT value

Figure 3.14 displays an example of a path derived from the network illustrated in Figure 3.13. This path consists of four edges (Examined Account→A, A→B, B→C, and C→Sink). For this path, three TT values are calculated, namely from the Examined Account to A ($TT_{ExaminedAccount\rightarrow A}$), from A to B ($TT_{A\rightarrow B}$), and finally from B to C ($TT_{B\rightarrow C}$). The total TT value of the whole path is the multiplication of these three calculated values and is assigned to the sink node. Figure 3.14 also presents the equation calculation the $TT_{A\rightarrow B}$.

3.6.2.2. Experimental Results

In this section, the TTT values of the created networks with respect to the examined accounts are presented. Table 3.3 is divided into four parts. Each part refers to the separate samplings mentioned above, while the related information per part is:

- the number of followers that is iteratively fetched,
- the number of top-k followers of the two generated categories; these are "by Influence Metric" and "by Followers", which are used for the creation of the respective layered networks,
- the distance threshold value that reflects the layers of the examined account's networks (TTL),
- the account which is the root of the two resulting networks (the six examined accounts),
- the TTT values of the two networks, according to the "by Influence Metric" and "by Followers", and, finally,
- the difference of the above TTT values for both generated networks.

Table 3.3: Details of each sampling set

Followers = 50, top-k users = 3, TTL = 3			
Username	By Influence	By Followers	Difference
@AdonisGeorgiadi	2,174	12,933	-10,759
@IliasKasidiaris	57,833	42,027	15,806
@PanosKammenos	22,527	30,074	-7,547
@SkaiGr	1,016	0,465	0,551
@YourAnonNews	0,038	0,018	0,020
@Pyrosvestiki	0,496	0,864	-0,368

Followers = 180, top-k users = 7, TTL = 3			
Username	By Influence	By Followers	Difference
@AdonisGeorgiadi	45,831	20,038	25,794
@IliasKasidiaris	682,280	592,961	89,319
@PanosKammenos	723,534	373,959	349,575
@SkaiGr	4,773	3,172	1,600
@YourAnonNews	2,234	0,980	1,255
@Pyrosvestiki	909,388	263,730	645,658

Followers = 100, top-k users = 5, TTL = 3			
Username	By Influence	By Followers	Difference
@AdonisGeorgiadi	12,733	8,632	4,102
@IliasKasidiaris	116,048	241,823	-125,775
@PanosKammenos	134,417	30,997	103,420
@SkaiGr	3,462	0,302	3,160
@YourAnonNews	1,762	0,446	1,316
@Pyrosvestiki	210,442	85,437	125,005

Followers = 360, top-k users = 7, TTL = 3			
Username	By Influence	By Followers	Difference
@AdonisGeorgiadi	50,686	15,503	35,183
@IliasKasidiaris	124,871	265,954	-141,083
@PanosKammenos	549,347	108,909	440,438
@SkaiGr	3,768	2,628	1,141
@YourAnonNews	3,844	2,866	0,978
@Pyrosvestiki	917,656	533,136	384,520

In addition, the green-highlighted values in column "Difference" correspond to the cases where the TTT value is larger in the "By Influence" category, thus indicating that our approach manages to create a network of followers who are more influential in comparison to the network of category "By Followers". The red-highlighted values indicate the opposite cases. As we can see, the wider the examined networks are in terms of the top-k accounts and their followers up to the third layer, the more influential network of accounts we have.

As can it be observed from Table 3.3, the TTT values of the two networks are raised as both the numbers of the followers and of the top-k accounts increase. The results of the use cases used for the evaluation of the influence metric calculation, show that the number of followers an account has is not solely sufficient to guarantee the maximum

diffusion of information in Twitter (and practically in any similar OSN). This is because, these followers should not only be active Twitter accounts, but they should also have an impact on the network. The latter is calculated by the Influence Metric value.

Chapter 4. Social Semantics

4.1. Introduction

On a daily basis hundreds of millions of messages are generated and disseminated by the numerous users of OSNs. This content is usually characterized by a highly unstructured and informal language, typographical errors, lack of structure, limited and often insufficient length, and high contextualization. Consequently, microblogging retrieval systems suffer from the problems of data sparseness and the semantic gap (Kalloubi et al., 2016).

To overcome these limitations and to contextualize the semantic meaning of the microblog content, recent research interest is focused not only on exploiting the existence of social semantics and user-generated content by identifying entities in them, but also on enriching and organizing this unstructured data. These entities can be utilized as keywords providing topical insights on the messages and consequently to their authors, description of real-time events, as well as revealing behavioral patterns utilized for building interest profiles. Often, those entities are linked to knowledge bases (e.g. Google Knowledge Graph), to the Linked Open Data (LOD) cloud (e.g. DBpedia) or they are represented as concepts extracted from ontologies using Semantic Web vocabularies in order to transform unstructured data into Linked Data.

In this chapter, we focus on the importance of social semantics in OSNs, and the limitations that derive from their diversity, and we propose novel and adaptable methodologies for improving the state-of-the-art. To this end, we present the "*InfluenceTracker Ontology*" (Figure 4.4) for transforming unstructured social data into Linked Data, and, specifically, for modeling Twitter accounts, their metadata, their social relationships, the entities included in their tweets (mentions, replies, hashtags, photos, URLs), as well as other social and qualitative metrics. As a use case, we apply this ontology on Twitter; its properties have been defined in such a way so as to be easily extensible to cover concepts from others OSNs as well. Our schema compared to the ones described in the related literature, is richer, more robust and able to represent a wider range of social information.

Moreover, having this ontological schema as a backbone, we present the architecture and infrastructure of "*InfluenceTracker*[8]", a publicly available website where anyone can measure the importance of any Twitter account, view and compare its activity and entities included in the tweets (mentions, replies, hashtags, photos, URLs), discover additional information about it, as well as perform sophisticated SPARQL queries. In addition, we describe the ontology used for transforming the Twitter accounts, their metadata, their social relationships, and the entities as well as other social metrics used in the Linked Data literature. In order to increase the data value and to enrich the semantified Twitter-related social information, we interlinked our social analytics with the LOD cloud, and, specifically, with the DBpedia knowledge base, thus fulfilling all the preconditions to characterize our data as a five star model, according to Tim Berners-Lee's rating system (Berners-Lee, 2012). Finally, in order to automatically discover these relationships, we propose and analyze two generic and adaptable methodologies for discovering the necessary linked data resources (URIs)

from external knowledge bases describing the Twitter entities in the best possible way.

To summarize, this chapter provides the following contributions:

1. We define the "*InfluenceTracker Ontology*[13]", an easily extensible schema, for modeling Twitter social analytics as Linked Data.
2. We present the architecture and infrastructure of "*InfluenceTracker*[8]", a publicly available service providing social analytics enriched by the LOD cloud.
3. We analyze two generic methodologies for linking the semantified Twitter entities with the LOD cloud.
4. We present the "*InfluenceTracker*" dataset, a five-star data model, which is a part of the LOD cloud since 20/02/2017.

The rest of this chapter is outlined as follows. In section 4.2 we discuss related work, while in section 4.3 we analytically present the architecture of the InfluenceTracker service, define the employed ontology behind our service for the transformation of the raw instructed data from Twitter API into Linked Data, and present a SPARQL endpoint where our five-star data can be queried. Finally, in section 4.4 we describe the methodology followed in order to discover which DBpedia URIs describe the requested Twitter entities in the best possible way.

4.2. The role of Semantics

As the adoption of semantics and Linked Data increases, many works have emerged covering aspects of semantic modeling in OSNs. In this section, we present approaches which adopt Semantic Web technologies for modeling the logical topology and structure of social networks and media as well as for transforming unstructured data into Linked Data.

One of the first studies in this domain is (Hepp, 2010), where the use of a specific syntax is proposed for creating a common knowledge representation, by incorporating RDF-like syntaxes into Twitter posts. The use of such statements enables users to freely define relations such as hierarchical or equality ones among hashtags.

Another work on the enrichment of Twitter messages with semantics is described in (Abel et al., 2011). The authors attempt to create user profiles by exploiting Twitter posts by using Semantic Web technologies. In order to capture the users' interests, the URLs of news articles found in tweets are utilized. A lexical analysis is applied on their content so that the relationships among the entities in news articles (representing the interests) can be discovered, which are then semantically related to those tweets.

The authors in (Wang et al., 2016) create graphs of hashtags found in tweets and utilize their relational information in order to discover latent word semantic connections in cases where words do not co-occur within a specific tweet. Sparseness and noise in tweets are handled by exploiting two types of hashtag relationships: i) explicit ones, which refer to hashtags that are contained in a tweet, and ii) potential ones, which refer to hashtags that do not appear in a tweet but co-occur with others.

Finally, the hashtags and words which have the highest probability to appear on a specific topic are discovered.

Another set of studies employ Semantic Web technologies, ontologies and the DBpedia knowledge base, which is a semantified version of Wikipedia, to achieve their goals.

The study in (Shinavier, 2010) introduces a semantic data aggregator in Twitter, which utilizes Semantic Web vocabularies in order to transform social data into structured microblog content. This framework focuses on the provision of Twitter messages as user-driven Linked Data, and, more specifically, metadata associated with the authors and the content of those social posts.

Another framework which utilizes semantic technologies, common vocabularies and Linked Data in order to extract microblogging data from scientific events from Twitter is proposed in (Vocht et al., 2011). In this work, the authors introduce a methodology for identifying similar users and organizations according to geospatial and topic entities.

The authors in (Slabbekoorn et al., 2016) propose an ontology-assisted topic modeling technique for determining the topical similarities among Twitter users. The entities found at the posts are mapped to classes of the DBpedia ontology, using the DBpedia Spotlight tool, and they are used for the labeling of clusters. Moreover, the topical similarities among individuals on different topics are calculated using ranking techniques, which define the structure of the resulting graphs. Based on these graphs, a quasi-clique community detection algorithm is applied for the discovery of topic clusters without predefining their target number.

Finally, a theoretical use of the benefits of the LOD is presented in (Shabir and Clarke, 2009), where the authors propose the use of such data for educational purposes.

Although ontologies and semantic technologies have been used in other works, neither of them captures and models such a wide range of information, spanning from the Twitter related characteristics of the accounts to the entities found in the posted messages, nor are easily expandable as to be applied to other OSNs, as the proposed InfluenceTracker ontological schema, which we described in Section 4.3.2.

A detailed presentation of studies based on Semantic Web technologies along with network theory and graph properties for transforming unstructured data into Linked Data, topic identification, detection of similar users and communities, as well as user personalization (e.g. interests, suggestions, and so on) can be found in Section 2.5.1.

4.3. The InfluenceTracker service

In this section, we present the architecture and infrastructure of the InfluenceTracker.com service, a publicly available website where anyone can rate and compare the recent activity of any Twitter account. In addition, we describe the ontology used for transforming the Twitter accounts, their metadata, their social relationships, the entities included in their tweets (mentions, replies, hashtags, photos, URLs), as well as other social metrics into an RDF graph.

4.3.1. Architecture

The architecture of the InfluenceTracker.com service and the relevant data flows are presented in Figure 4.1. The service combines the use of a relational database joint with an RDF triple store. Thus, the data and the related information displayed at the web pages combine both technologies. The relational database is a MySQL Server and the RDF triple store is contained in an Open Link Virtuoso (OLV) Server. As it can be seen, our service is related to the DBpedia knowledge base, thus the presented data may also be enriched from that source. There are three use case scenarios of the service.

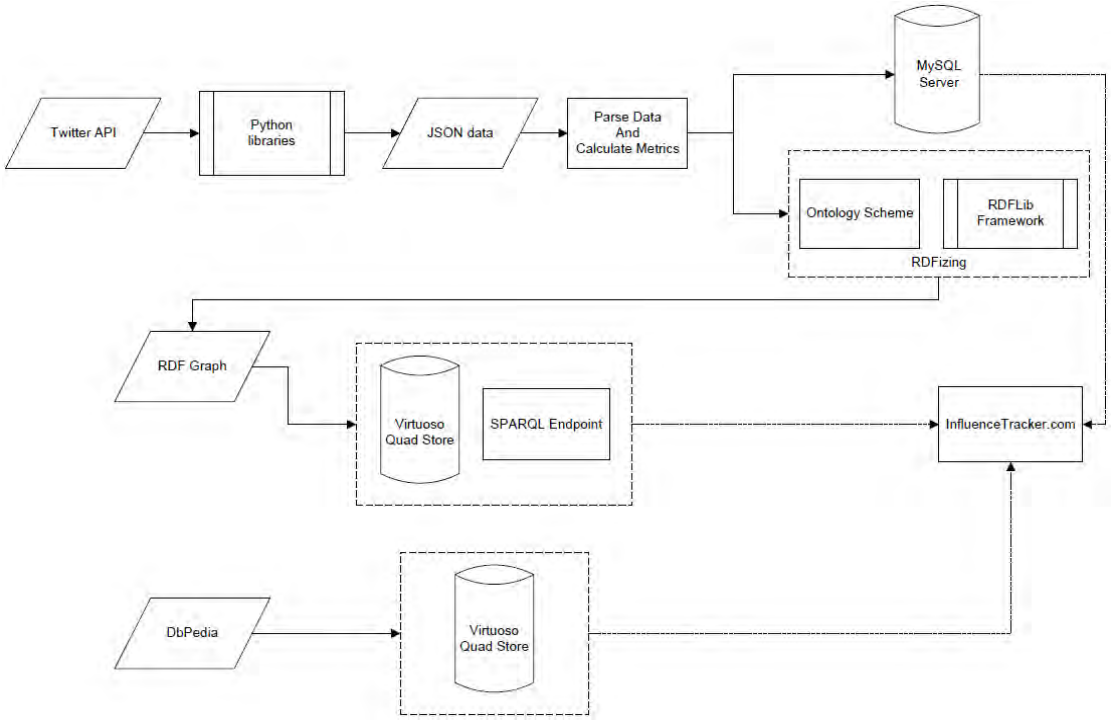

Figure 4.1: The architecture of the InfluenceTracker.com service

The first of these scenarios involves the update of the RDF graph. An implemented service based on Python libraries, is executed on a regular basis. The process is split into four phases. During the first phase, a request is sent to the Twitter API for each account found in the database. The response contains the data in JSON format. In the second phase, the necessary data are parsed and the metrics (some of them were presented in Chapter 3) are calculated. The third phase involves the semantification of the collected data with concepts (resources and property URIs) derived from our ontology (see Figure 4.4 and Section 4.3.2) and from the RDF graph updates. This process is performed by using the RDFLib framework. During the last phase, the triples are stored in the OLV environment, while the user can use a SPARQL endpoint[10,11] for custom semantic searches.

The second use case is a subset of the previous one. It takes place when a Twitter account is searched through the provided web interface[8]. Another service, also

[10] http://www.influencetracker.com/endpoint
[11] http://www.influencetracker.com:8890/sparql

implemented in Python, performs a request to the Twitter API for the investigated account. A response is returned in JSON format. In case of a valid account, the necessary data are parsed, the related metrics are calculated and stored in the relational database. In this use case, no data are stored in the RDF graph. This is because we wanted to maximize the responsiveness of our service, minimizing in parallel the execution time. Finally, in case where a new account is inserted into the system, the necessary data will be stored at the RDF graph during the next update process.

The third scenario involves the use of the service. When an account is selected for display[12], an informative table and a set of historical data, as recorded by the system, are presented. In case that the Twitter account is linked with a DBpedia URI (see Section 4.4), a request is made to that service using semantic technologies in order to retrieve and display the necessary information. This table also contains social analytics data. Depending on the type of the described entity, different kinds of information are presented. Figure 4.2 and Figure 4.3 present such examples of two different types of entities, a broadcaster and a person respectively. The historical data are retrieved by both the RDF graph and the MySQL Server.

	Cable News Network - CNN	Retrieved on: 04/07/2015
	The Cable News Network (CNN) is an American basic cable and satellite television channel that is owned by the Turner Broadcasting System division of Time Warner. The 24-hour cable news channel was founded in 1980 by American media proprietor Ted Turner.	
	It's our job to #GoThere and tell the most difficult stories. Come with us!	
Twitter Details		**DBpedia Details**	
Influence Metric	73.04	Category	Broadcaster, TV Station
Tweets	63,744	Country	United States
Tweets Per Day	49.45	Broadcast Area	Canada, United States
ReTweet Percentage	44.0 %	Slogan	Go There, This is CNN
Reply Percentage	1.0 %	Launch	1980-06-01
Followers	18,052,113	Homepage	http://www.cnn.com/
Following	1,100	Wikipedia Page	http://en.wikipedia.org/wiki/CNN
ReTweet h-index - Last 100 Tweets	52	DBpedia URI	http://dbpedia.org/resource/CNN
Favorite h-index - Last 100 Tweets	54	Times Searched	3

Figure 4.2: The informative table of Twitter account "CNN" (@cnn) combining social analytics and DBpedia data

[12] http://www.influencetracker.com/searchedAccounts

Berners-Lee, Tim, Sir Tim Berners-Lee, Tim Berners-Lee - Tim Berners-Lee
Sir Timothy John "Tim" Berners-Lee, OM, KBE, FRS, FREng, FRSA, DFBCS (born 8 June 1955), also known as "TimBL", is a British computer scientist, best known as the inventor of the World Wide Web.
Director of the World Wide Web Consortium (W3C) w3.org, the place to agree on web standards. Founded webfoundation.org - let the web serve humanity

Twitter Details		DBpedia Details	
Influence Metric	44.48	Category	Person
Tweets	631	Birth Date	1955-06-08+02:00
Tweets Per Day	0.21	Birth Place	London, England, United Kingdom
ReTweet Percentage	57.0 %	Short description	British computer scientist, best known as the inventor of the World Wide Web, British computer scientist, best known as the inventor of the World Wide Web
Reply Percentage	11.0 %	Homepage	-
Followers	219,921	Wikipedia Page	http://en.wikipedia.org/wiki/Tim_Berners-Lee
Following	332	DBpedia URI	http://dbpedia.org/resource/Tim_Berners-Lee
ReTweet h-index - Last 100 Tweets	24	Times Searched	4
Favorite h-index - Last 100 Tweets	23	Retrieved on	25/10/2015

Figure 4.3: The informative table of Twitter account "Tim Berners-Lee" (@timberners_lee) combining social analytics and DBpedia data

4.3.2. The InfluenceTracker Ontology

The proposed InfluenceTracker ontology[13] utilizes classes and properties from the FOAF[14] ontology (Brickley and Miller, 2004). FOAF (Friend-of-a-Friend) is an ontology for describing persons and their activities as well as their relations to other people and objects, while it can be generalized as to describe all types of entities, and called agents, who are responsible for specific actions (Brickley and Miller, 2004). In our context the agents are the Twitter users, who are responsible for specific actions, such as owing Twitter accounts, posting tweets, and interacting with others. Figure 4.4 displays the classes and their hierarchical relationships, where highlighted the FOAF ontology classes are. An OWL version of the ontology is also available[13].

During the representation of the entities, two specific prefixes are used, namely "foaf" and "it". They correspond respectively to the namespaces of the FOAF and of the proposed ontology. The ontology is built on three basic building blocks, namely classes, as well as object and datatype properties.

[13] http://www.influencetracker.com/ontology
[14] http://xmlns.com/foaf/spec/

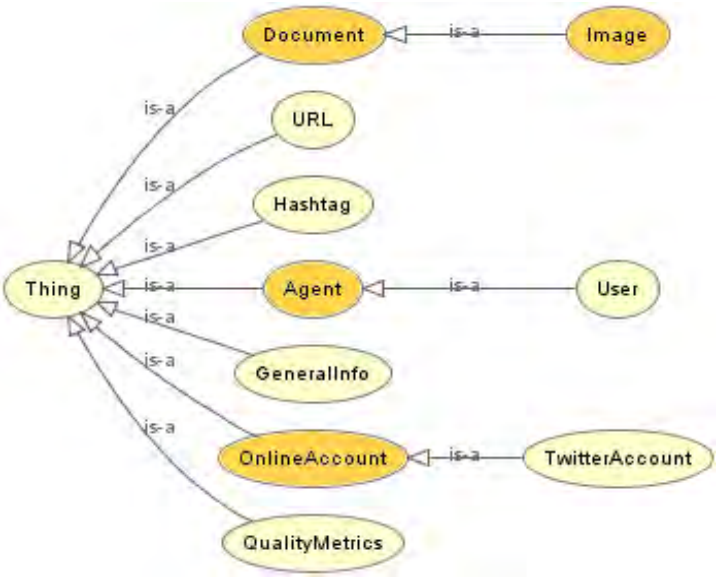

Figure 4.4: The hierarchy of the classes of the "InfluenceTracker" ontology

4.3.2.1. Classes

The classes are used to represent conceptual entities. The ones defined in the InfluenceTracker ontology are the following:

- foaf:Agent: A general class which describes agents who are responsible for several actions.
- it:User: A subclass of the foaf:Agent, which describes the agents that own a Twitter account. These may be physical persons, organizations, events, parties etc.
- foaf:OnlineAccount: It represents the provision of some form of online service, by some party (indicated indirectly via the foaf:accountServiceHomepage object property) to some foaf:Agent.
- it:TwitterAccount: A subclass of the foaf:OnlineAccount, representing the actual Twitter accounts.
- it:GeneralInfo: This class contains the Twitter related details of an account characterized by InfluenceTracker.com as "General Information". These are the total number of tweets, the *TCR*, the retweet ratio, and the number of followers and following.
- it:QualityMetrics: This class contains the metrics of a Twitter account characterized by InfluenceTracker.com as "Quality Metrics". These are the *"Retweet and Favorite h-index - Last 100 Tweets"*, the estimated *"Retweet and Favorite h-index"*, the reply ratio and the value of our influence metric.
- foaf:Document: This class represents those things which are broadly conceived documents. There is no distinction between physical and electronic ones.

- foaf:Image: This class corresponds to the image documents. It is a subclass of the foaf:Document, since all the images are documents. Digital images are instances of this class.
- it:Hashtag: The class describes the entities which are hashtags (words starting with "#").
- it:URL: The class describes the entities which are URLs.

4.3.2.2. Object Properties

The object properties are those for which their value is an individual. The ones defined in the InfluenceTracker ontology along with their concept restrictions are the following:

- foaf:account: This property is used to relate a foaf:Agent to a foaf:OnlineAccount for which it is the sole account holder.
- foaf:accountServiceHomepage: This property indicates a relationship between a foaf:OnlineAccount and the homepage of the supporting service provider.
- it:hasGeneralInfo: This property relates an it:User to an it:GeneralInfo which contains the Twitter related information of the owned account, characterized by InfluenceTracker.com as "General Information".
- it:hasMentioned: This property relates an it:User to an it:User that has been mentioned in the first user's tweets.
- it:hasQualityMetrics: This property relates an it:User to an it:QualityMetrics which contains the metrics of the owned account, characterized by InfluenceTracker.com as "Quality Metrics".
- it:hasRepliedTo: This property relates an it:User to an it:User that has received a tweet as a reply from the first user.
- it:includedHashtag: This property relates an it:User to an it:Hashtag that has been included in the user's tweets.
- it:includedImage: This property relates an it:User to an it:Image that has been included in the user's tweets.
- it:includedUrl: This property relates an it:User to an it:URL that has been included in the user's tweets.
- it:isFollowing: This property relates an it:TwitterAccount to an it:TwitterAccount in cases where the first account follows the second one. It represents the action called "Follow" introduced by Twitter. It is the reverse property of it:hasFollower.
- it:hasFollower: This property relates an it:TwitterAccount to an it:TwitterAccount in cases where the second account is a follower of the first one. It is the reverse property of it:isFollowing.
- it:hasSimilar: This property relates an it:TwitterAccount to an it:TwitterAccount in case that they are characterized as similar in terms of their disseminated content and Twitter entities used (see Section 5.4).

- it:dbpediaUri: This property relates an it:TwitterAccount to its respective DBpedia URI (see Section 4.4).

These properties have been defined in such a way so as to be easily extensible to cover concepts from other OSNs as well. If the Twitter accounts are replaced by those of Facebook, then the tweets are the statuses. The actions of "Share" and "Like" found in Facebook are the equivalent of "Retweet" and "Favorite" of Twitter. The concepts of hashtags, mentions, replies, images and URLs are practically the same in both of these OSNs.

4.3.2.3. Datatype Properties

The datatype properties are those for which their value is a data literal. Those defined in the InfluenceTracker ontology along with their concept restrictions are the following:

- foaf:accountName: This property provides a textual representation of the account name (unique ID) associated with that account.
- it:description: This property provides the description of an account, as set by its owner.
- it:displayName: This property provides the name displayed at the web page of an account, as set by its owner.
- it:followers: This property provides the number of the followers of an account.
- it:following: This property provides the number of the accounts that an account follows.
- it:hIndexFav: This property provides the value of the *"Favorite h-index - Last 100 Tweets"* metric of an account.
- it:hIndexFavDaily: This property provides the estimated daily value of the *"Favorite h-index"* metric during the lifespan of an account.
- it:hIndexRt: This property provides the value of the *"Retweet h-index - Last 100 Tweets"* metric of an account.
- it:hIndexRtDaily: This property provides the estimated daily value of the *"Retweet h-index"* metric during the lifespan of an account.
- it:imageUrl: This property provides the URL that leads to an image which was included in a tweet.
- it:influenceMetric: This property provides the value of the Influence Metric measurement (see Section 3.4). Its aim is to describe both the importance and the impact of an account in a social network.
- it:profileLocked: This property indicates whether the profile of an account is publicly visible or not.
- it:activeAccount: This property indicates whether an account is active or not.

- it:replyRatio: This property provides the ratio of the user's latest tweets which are used as replies to other users' tweets.
- it:retrievedOn: This property provides the date that the information regarding an account was lastly updated.
- it:rtPercent: This property provides the percentage of the latest user's tweets that are retweets from other accounts.
- it:tweets: This property provides the number of the total tweets posted by an account.
- it:tweetsPerDay: This property provides the number of the average tweets posted per day by an account.
- it:url: This property provides the short URL that leads to a web site which was included in a tweet.
- it:fullUrl: This property provides the full URL representation of a shortened one which was included in a tweet.
- it:domain: This property provides the domain of a URL which was included in a tweet.

4.3.3. Federated SPARQL Queries

As already mentioned, a public endpoint allows the search of the collected semantic data along with their combinations and enrichment with others from publicly available datasets from the LOD cloud. The federated SPARQL Query 1 (in Appendix C: SPARQL Queries) returns some Twitter related information of an account (i.e. the displayed name, and the values of Influence Metric and "ReTweet h-index - Last 100 Tweets"), its related DBpedia URI and some DBpedia related information (i.e. the person's birth date, birth place and short description). The federated queries demonstrate the benefits of the semantic technologies and the LOD cloud, since they are able to provide answers to sophisticated queries (e.g. return the top-10 members of political parties according to their Influence Metric value).

As already mentioned, a public endpoint[10,11] allows the search of the collected semantic data. The URIs which are returned by the queries are dereferenceable ones, consequently the resources that they identify are represented by documents, which in our case are in HTML format. These URIs are constructed using the Slash format. An example of such a URI is "http://www.influencetracker.com/resource/User/youtube". It represents the document where the resource "youtube", an instance of the "it:User" class, is described. An instance of that document can be seen in Figure 4.5.

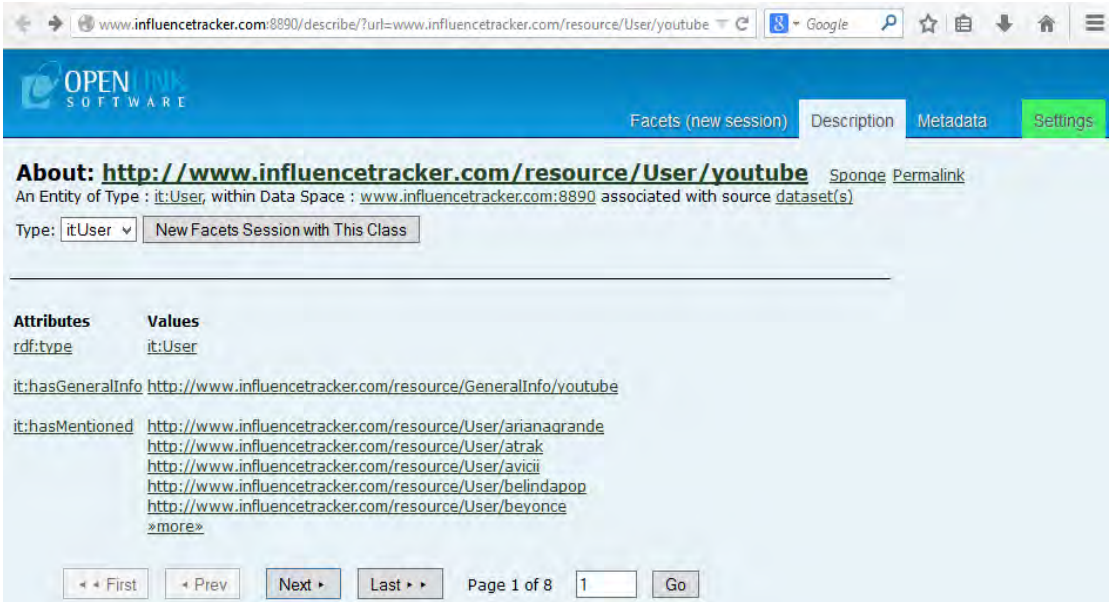

Figure 4.5: The document in HTML format of a dereferenceable URI

Publishing and consuming LOD lead to the enrichment of the existing information and to the increase of their value, as new data of interest are discovered (Bauer and Kaltenböck, 2012). As seen from the aforementioned federated Query 1 (in Appendix C: SPARQL Queries), the semantic technologies allow us to treat multiple distinct datasets as one. Therefore, the Twitter related data existing in the InfluenceTracker (IT) graph can be expanded and combined using information from the DBpedia graph and vice-versa.

Despite each dataset having been modeled according to its proposed ontological specification, the combination of the data allows us to also handle different ontological specifications as one. In our case, a combined view of the involved ontologies can be found in Figure 4.6. The highlighted items are the classes of the IT ontology, while the rest belong to the DBpedia (Lehmann et al., 2015) one which incorporates classes and properties from other well-established ontological schemes, such as FOAF[14], YAGO[15], and schema.org. As it can be seen, the foaf:Agent class is used by both ontologies and is further expanded to cover more concepts (e.g. persons, organizations). It should be noted that for simplicity reasons only a portion of the combined ontology is presented; specifically only the classes with the highest number of instances combined with our data.

By interlinking the InfluenceTracker semantified data with resources from DBpedia, we fulfilled all the preconditions to characterize our data as a five-star data model, according to the Tim Berners-Lee's LOD rating system (Berners-Lee, 2012). Thus, since the latest update of the LOD cloud[16], on 20/02/2017, the "*InfluenceTracker*" dataset is officially part of this interlinked and interdependent ecosystem of data. This dataset[17] along with other useful information are publicly available at the DataHub[18] open data repository.

[15] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/#c10444
[16] http://lod-cloud.net/versions/2017-02-20/lod.svg
[17] https://old.datahub.io/dataset/influence-tracker-dataset

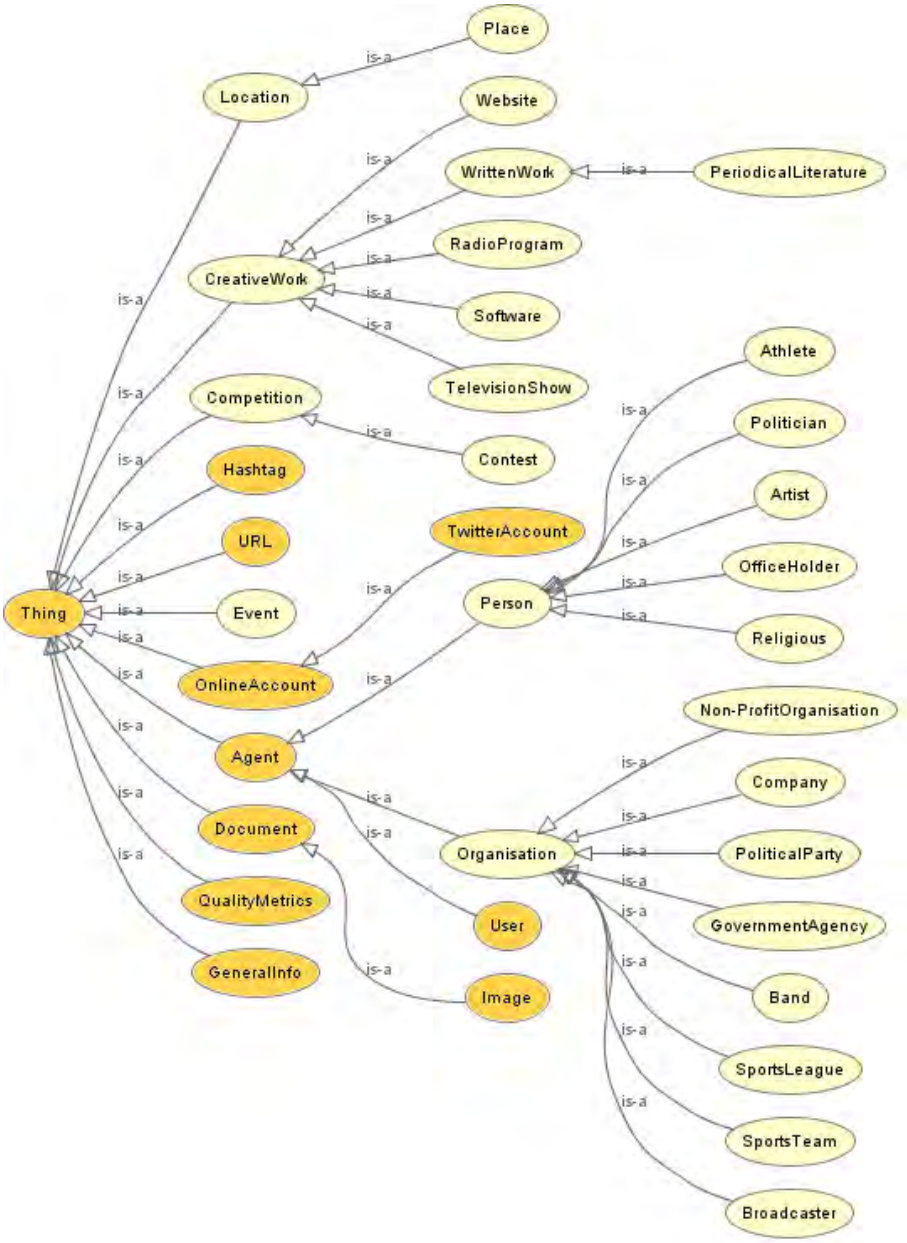

Figure 4.6: A portion of the combined view of the InfluenceTracker and DBpedia ontologies

4.4. Discovering DBpedia URIs

One of the issues faced by the publishers of LOD is the effort that has to be put in order to link the data to other sources found on the web (Michelson and Macskassy, 2010). This was one of the main impediments that we also had to overcome. The URIs found in the IT graph had to be linked with the appropriate DBpedia ones, or simply put to discover which DBpedia URI describes best a Twitter account.

In order to discover these URIs a certain methodology was employed. This methodology consists of two parts. In the first part a DBpedia URI is searched

[18] https://datahub.io/

according to the “Display Name” of a Twitter account, as provided by its creator or user. The process is as follows:

1) retrieve the “Display Name” of the URIs of the Twitter accounts found in the IT graph, and then
2) search in the DBpedia graph for the URI which has as name the provided “Display Name”. Three properties are used for modeling the “name” attribute: foaf:name, dbprop:name, and rdfs:label.

If the value of any of the three properties matches the “Display Name” then that DBpedia URI is likely to be correct and it is stored.

If a URI is not found, then the second part of the methodology is initiated. Its aim is to attempt to create the DBpedia URI by formatting the “Display Name” under certain name pattern conventions, such as the ones followed by DBpedia. These conventions are the following:

- the white-spaces are replaced by underscores,
- the first letter of each word is capitalized,
- Greek and other special characters from other alphabets are transliterated according to the DBpedia conventions (e.g. “é” to “e”, “ü” to “u”, “ψ” to “ps”),
- in case of physical persons, the last fragment of the URI is their name in the “First Name - Last Name” format.

Such an example is the “Display Name” with the value “Φώτης Κουβέλης”. According to the aforementioned conventions it is transformed into “Fotis_Kouvelis”, which results in a valid DBpedia URI.

Even after these steps no URI is found, a final transformation is applied. This transformation includes the swapping in the order of the words of the transliterated “Display Name”. This covers the cases where the URI is not provided in the necessary by the DBpedia format “First Name - Last Name”.

During the second part of the methodology, even if a URI is found it may not be the appropriate one and further handling or review may be necessary.

URIs needing further handling are those used for redirection to others (e.g. http://dbpedia.org/resource/NBA redirects to http://dbpedia.org/resource/National_Basketball_Association) or for disambiguation among different entities, usually having the same name (e.g. http://dbpedia.org/resource/Android).

The vast amounts of data found in the DBpedia graph and the lack of a standard input by the users of the Twitter accounts lead in being almost a necessity to further review and evaluate the derived URIs. An example of such a scenario is the case where the “Display Name” “Honda” is searched. The application of the aforementioned methodology leads in a match from the DBpedia graph, but it was referring to a comet and not to the well-known company. The name of the latter in the DBpedia graph is “Honda Motor Co., Ltd.”. Other occasions of wrongly matched URIs are cases of

synonymity among various entities, a phenomenon usually occurring among natural persons.

At the end of the process the derived DBpedia URIs are linked with the respective ones of the IT graph for further use.

In order to measure the effectiveness of the two aforementioned methodologies we calculated their precision and recall. As it can be seen in Table 4.1, the "Search by Name" methodology is outperformed by the "URI Construction" one. The precision of the first is 64.4% while the precision of the latter is significantly higher at 87.7%, while their recall is 38.42% and 40.39% respectively. Despite the fact the latter methodology results in less URIs, the true positives are higher while the false ones are substantially lower, leading to an increased precision and recall of 36% and 5% respectively.

Table 4.1: Effectiveness of the URI Discovery Phases

Phase	Elements	Counter	Precision	Recall
Search by Name	True Positives	156	64.46%	38.42%
	False Positives	86		
	Total retrieved	242		
URI Construction	True Positives	164	87.7%	40.39%
	False Positives	23		
	Total retrieved	187		
Total DBpedia URIs			406	

In order to further investigate the effectiveness of the two methodologies we divided the DBpedia resources derived by the proposed framework into 12 categories. For each of them we calculated its precision and the percentage of the false positives. The results showed that there are cases where the two methodologies perform approximately the same but in others the "URI Construction" performs much better. The results for the categories of "Natural Persons", "Companies" and "Press" are presented in Table 4.2 (in Appendix B: DBpedia Categories). As presented in Table 4.1, the number of the derived URIs of the "URI Construction" methodology is lower, but the number of true positives is higher and the false ones are substantially lower.

Chapter 5. Social Identification

5.1. Introduction

Twitter is an Online Social Network (OSN) where millions of accounts publish their messages on a daily basis. Despite the fact that these accounts represent numerous types of agents (e.g. persons, companies, landmarks, parties) and each one has a different impact on the network, they all share common interests on specific thematic categories. These can be exploited for revealing behavioral patterns and for building interest profiles, thus enabling the interrelation of semantically related terms and the social proximity or similarity between profiles and interests.

In this context, we present a framework towards the discovery and suggestion of similar accounts in Twitter, in terms of their disseminated social entities (mentions, replies, hashtags, photos, and URLs). This framework is based on the assumption that the more common entities are found in the disseminated messages of OSN accounts, the more similar, in terms of content or interest, they tend to be. The proposed "*Similarity Metric*" is calculated using exclusively semantic mechanisms and technologies and utilizing the structure defined in the *InfluenceTracker Ontology*[13] (Section 4.3.2).

Moreover, on top of this methodology, we developed an iterative algorithm towards the automatic labeling with respect to thematic categories derived from properties of the DBpedia knowledge base, in order to classify them into communities. The "Thematic Category Labeling Algorithm" demonstrates the benefits of the semantic technologies and the LOD cloud, thus increasing the value of the available data. The enriched social information is able to provide answers to sophisticated queries, such as: return the top-10 members of political parties according to their "Influence Metric" value who have used hashtag "x". To the best of our knowledge, there is currently no active service for providing such kind of data linkage, i.e. social analytics with the LOD cloud.

To summarize, this chapter provides the following contributions:

1. We present the "*Similarity Metric*", a framework for discovering and suggesting similar accounts in Twitter, in terms of their disseminated social entities, based on semantic mechanisms and technologies.
2. We propose the "*Thematic Category Labeling Algorithm*" towards the automatic labeling and classification of Twitter accounts with respect to thematic categories derived from the DBpedia knowledge base.
3. We discuss the results of the case study, and we further evaluate our methodology against subjective ratings from 22 evaluators.

The rest of this chapter is outlined as follows. In sections 5.2 and 5.3 we discuss related work on social interest identification and social recommendation. In section 5.4 we present our approach towards similarity recommendation for Twitter accounts. In order to gain insight into this methodology we analytically describe a case study. In section 5.5 we describe the rationale behind the selection of these thematic categories derived from DBpedia properties, and we discuss their evaluation assessment.

Furthermore, we analytically present the proposed algorithm towards the automatic labeling of Twitter accounts with respect to the thematic categories and we present an overview of the results. Finally, in section 5.6, we evaluate and discuss the results of the case study, and we further evaluate our methodology against the subjective ratings from 22 evaluators.

5.2. Social Interests

Despite the fact that the set of social semantics of each account in an OSN is unique, as they depend on personal social activities, common patterns among them can be recognized. These can be exploited to enable the discovery of the users' social behavior interests, and preferences.

The authors in (Michelson and Macskassy, 2010) present a topic-oriented framework for Twitter, aiming at discovering the users' topics of interest by examining the entities found in their posts, which may be mentions or plain text (in OSNs the mentions are words prefixed with "@"). The Wikipedia knowledge base is leveraged in order to disambiguate those entities and the topics of interest to be defined (e.g. the term "apple" may refer to the fruit or to the multinational technology company).

Topic profiling using the Wikipedia knowledge base is also studied in (Lim and Datta, 2013). The topics are discovered based on the hashtags posted by Twitter users and their friends. The celebrities (accounts of popular people) who are followed are the primary indicators of interest which have been derived from their Wikipedia classification. The indicators along with the hashtags infer the topics of interest of the accounts.

Similarly, the work in (Kapanipathi et al., 2014) focused on extracting the interests of Twitter accounts based on their generated messages. The methodology leverages the hierarchical relationships found in Wikipedia in order to infer user interests. The authors claim that the hierarchical structures can improve the existing systems to become more personalized based on broader and higher level concepts (e.g. the concept "Basketball" is more generic that the term "NBA").

A topic-oriented framework, which uses the DBpedia knowledge base, is proposed in (Kalloubi et al., 2016). Specifically, the context of Twitter is mapped to DBpedia entities and graph-based centrality theory is applied for assigning weights to the entities of the examined messages.

A framework for discovering similar accounts in Twitter based only on the "List" feature is proposed in (Kanungsukkasem and Leelanupab, 2016). This functionality allows the users to create their own lists by adding any account they wish. The authors claim that this feature is considered a form of crowd-sourcing. The hypothesis of the methodology is that when two accounts are present in the same list they should be similar or related to each other. Therefore, the proposed metric relies on the number of lists that a specified account and a potentially similar one are listed together.

Another framework employed for inferring user interests in Twitter is presented in (Besel et al., 2016). Contrary to the previous frameworks, it is based on the users' followees and the content they consume, rather than on their original posts. This proposal is based on the hypothesis that famous people maintain accounts being followed by a large number of users. The Wikipedia articles of the former are

discovered, linking to a higher level of categories and hierarchies, which become an implicit expression of the users' interests.

The following studies employ semantic technologies and related protocols that can be utilized to represent user preferences and similarities. In (Blei et al., 2003) the authors employ Semantic Web technologies for the creation of user profiles by analyzing Twitter posts. In order to capture the users' interests, the URLs of news articles found in tweets are exploited. Lexical analysis is applied on their content in order to discover the relationships between the entities in news articles (representing the interests) which are then semantically related to these tweets. The framework in (Vocht et al., 2011) exploits common vocabularies and Linked Data in order to extract microblogging data regarding scientific events from Twitter. Finally, an ontology-assisted topic modeling technique for determining the topical similarities among Twitter users is proposed in (Slabbekoorn et al., 2016). The entities found at the posts are mapped to classes of the DBpedia ontology and are used for the labeling of clusters. Moreover, the topical similarities among individuals on different topics are calculated using ranking techniques, which define the structure of the resulting graphs. Based on these graphs, a quasi-clique community detection algorithm is applied for the discovery of topic clusters, without predefining their target number.

In Section 2.5.2 the interested reader can find a wider coverage of studies employing social semantics for identifying similar properties and activities with respect to user-generated content, description of real-life events, as well as revealing user interests and behavioral patterns across different online social media users.

5.3. Social Recommendation

The studies presented in this section describe approaches on recommendation systems which utilize the available information in OSNs for proposing social content or accounts based on the users' profiles. An interesting problem in the area of social recommendation systems is the suggestion of which similar users to follow. As the latter share common attributes and characteristics, they can be grouped into potentially overlapping communities providing useful insights not only for the analysis of OSNs but also for understanding the structure and the properties of complex networks.

A semantic followee recommender system in Twitter is proposed in (Deb et al., 2016). This system integrates content-based filtering approaches based on Twitter, and popularity identification among users using collaborative-filtering over the friendship network, along with publicly available knowledge resources (i.e. Wikipedia, WordNet, Google corpus). The aim is to classify the tweets into six classes and to label the users as a recommendation service. The application of a Kalman filter enables noise removal and the prediction of future tweet patterns leading to a new multi-labeling of the users. A ranking-based followee recommendation scheme in microblogging systems that is based on the latent factor model is proposed in (Chen et al., 2016). To model user preferences both tweet content (original posts and retweets) and social relation information (followers, followees) are taken into consideration. Another followee recommendation methodology that builds interest profiles is proposed in (Hannon et al., 2010). These profiles are built by exploiting not only the users' generated content but also of their directly related ones (followers, followees).

In (Ma et al., 2011), a matrix factorization framework with social regularization is proposed for improving the accuracy of recommender systems, by incorporating social network information. Social regularization includes two models for representing social constraints is based on the users'-friends' similarity at an individual and average level. Each social link is then weighted in accordance with the similarity among the users, allowing the exploitation of friends based on the rating similarity.

The LOD cloud is considered as a source for proposing recommendations. Such a study is presented in (Noia et al., 2012), where the authors utilize data from the LOD cloud in order to develop a content-based recommender movie system. Another work on a recommendation system based on the Linked Data and in particular on DBpedia is presented in (Passant, 2010). This system, called "dbrec", offers music recommendations. The author also describes how semantic distance measures can be applied to Linked Data, and proposes an ontology for representing such measures.

Considerable effort has also been devoted to recommendation systems for suggesting personalized streams of information ((Phelan et al., 2011), (Chen et al., 2010), (Tang et al., 2016), and (Zhang et al., 2016)).

"Buzzer" (Phelan et al., 2011) is such kind of a service for proposing news articles to Twitter users. To achieve that, terms from both the users' and their friends' timelines are mined. These terms act as ratings for promoting and filtering news content. The methodology described in (Zhang et al., 2016) is based on the same principles but it also incorporates additional factors affecting the interest of a user on a tweet, such as its quality, number of retweets, and the importance of its publisher.

URLs as a recommendation parameter in Twitter are examined in (Chen et al., 2010), with the scope to direct the users' attention through personalized suggestion mechanisms in Twitter posts. It is suggested that three main pillars should be considered in such kind of recommender, namely, the sources of the URLs, the users' area of interest, and social information.

Apart from the aforementioned studies, there are also other publicly available or commercial services for recommending Twitter accounts to be followed. Two of them, are the "Mofollow"[19] and the "Tweepi"[20]. The recommendations of the first service are based entirely on a user's friends' connections. "Tweepi" recommendations derive from contextual and relational aspects, in terms of posting messages containing predefined hashtags and of following predefined accounts. Even the suggestions[21] of Twitter are mainly based on the users' contacts, e-mails, locations, followers and followees. Minimal attention has been given to the content itself.

Most of the existing similarity identification approaches are either based on only a fraction of the available contextual information (e.g. -predefined- hashtags or URLs) or exclusively on friendship relations. Our work differentiates in terms of the extent of the social information we exploit, as well as in the employed technologies. In order to expand the data coverage, we utilize the entire available context, in terms of the disseminated social entities (mentions, replies, hashtags, photos, and URLs).

[19] http://www.mofollow.com/
[20] https://tweepi.com/how-it-works
[21] https://support.twitter.com/articles/227220-how-to-use-twitter-s-suggestions-for-who-to-follow

Moreover, our framework is based on exclusively semantic mechanisms and technologies and utilizes the *InfluenceTracker Ontology*[13] (Section 4.3.2).

The context-based identification of users' interests and similarities, along with the topological and structural attributes of the social networks can be used towards the identification of communities. An approach using node similarity techniques for community detection in OSNs is presented in (AlFalahi et al., 2013). A virtual network is created, where virtual edges are inserted based on the similarity of the nodes. The network is calculated using the Jaccard Measure. The proposed algorithm is then applied on the generated virtual network. Similarly, the study (Slabbekoorn et al., 2016) proposes a topic modeling technique among Twitter users using the DBpedia ontology.

In contrast to the traditional techniques, the approach in (Yang et al., 2013) uses not only the network topology but also the attributes of the nodes for developing a methodology towards the detection of overlapping communities.

Another algorithm for detecting communities is presented in (Qi et al., 2012). This algorithm is based on the content of the edges deriving from the users' pair wise interactions. According to the authors, this algorithm delivers a better perception of the communities because it depicts more effectively the nature of social interactions.

Most community detection algorithms use the edges of the nodes, their attributes or the structure of the graph in order to discover these communities. Our work differentiates in terms of the approach we follow. We utilize the results of our previous work (Razis and Anagnostopoulos, 2016) which produces weighted graphs of similar accounts, creating small communities, of suchlike attributes, interests and properties. That algorithm (Section 5.5) is iteratively applied and labels the accounts with the thematic categories derived from DBpedia, thus leading to the final classification.

5.4. Similarity Recommendation in Twitter

As already mentioned in Section 4.3, the InfluenceTracker.com service retrieves the Twitter accounts and their related social information in tweets, named entities (e.g. mentions, replies, URLs, hashtags, photographs) and stores them in an RDF graph. Common properties and patterns can be recognized among these accounts and their disseminated entities, which can be further exploited towards the discovery of the accounts' social behavioral patterns, interests, and preferences. Obviously, many of these entities are found in many tweets posted by different Twitter accounts. The more entities the accounts have used in common, the more similar their content tends to become. In order to measure and quantify the similarity of the Twitter accounts a methodology is proposed, which is presented below.

As suggested in (Naveed et al., 2011a), the presence of hashtags, mentions and URLs is typical in a tweet, thus they were utilized in their content-based framework. For the calculation of our Similarity Metric, four entities are used as comparison coefficients: the three "typical" ones (i.e. hashtags, mentions, and URLs) and additionally the domains of those URLs that an account has included in its tweets. The proposed methodology consists of the following seven steps:

1. define *k*, that is the number of the top similar accounts to be discovered,

2. define the *depth* of the similar accounts to be discovered (e.g. if depth equals to "two", then the top-k similar accounts of the top-k ones of the examined account will be iteratively searched and so forth),
3. define the account to discover its top-k similar ones,
4. retrieve the entities of an entity category of the examined account (e.g. all the hashtags included in its tweets),
5. discover all the Twitter accounts that included those entities in their tweets,
6. for each Twitter account find its total entity counter of the specific category (E_N) (e.g. how many hashtags have been tweeted),
7. find the common number of the specific category (E_{CN}) of the examined account with respect to others (e.g. their common number of hashtags),
8. calculate the coefficient of that specific category of Twitter entity (E_{Cf}) of the examined and each one of the other accounts (e.g. hashtag coefficient),
9. repeat steps 4 to 8 for the remaining entity categories,
10. depending on the depth value, repeat steps 3 to 9.

The coefficient of a specific entity category (E_{Cf}) is defined as the fraction of the common amount (counter) of the category (E_{CN}) by its total entity counter (E_N). The calculation of this coefficient is presented in Equation 5.1:

$$E_{Cf} = \frac{E_{CN}}{E_N}, \text{ where: } 0 \leq E_{Cf} \leq 1, E_{Cf} \in \mathcal{R}, E_N > 0, E_{CN} \geq 0. \qquad (5.1)$$

After step 7, four coefficients are calculated, namely "mention", "hashtag", "URL" and "domain" coefficient.

The next step is to utilize the resulting coefficients in order to calculate the Similarity Metric. Apart from the aforementioned coefficients, there are three other factors that are considered for the calculation of the Similarity Metric.

The first is the frequency of use of each of the four entity categories by the examined user, called "Entity Weight" (E_W). E_W is defined as the fraction of the entity counter of a specific entity category (E_N) by the sum of the entity counters of all entity categories (E_{SN}) and it is defined by Equation 5.2. Four weights are calculated -one for each entity category- namely "mention", "hashtag", "URL" and "domain" weight:

$$E_W = \frac{E_N}{E_{SN}}, \text{ where: } 0 \leq E_W \leq 1, E_W \in \mathcal{R}, E_N > 0, E_{SN} > 0. \qquad (5.2)$$

This factor is useful in cases where the E_{Cf} of an entity category of a compared account is high and the E_W of the examined account is significantly low. Moreover, the E_W is used for properly adjusting outlier E_{Cf} values. The resulting Weighted Coefficient (WC) of a specific entity category is defined in Equation 5.3:

$$E_{WC} = E_{Cf} * E_{W}. \qquad (5.3)$$

However, there are cases where the E_{WC} coefficient is not sufficient. Such a case is when two users have the same E_{Cf} value for an entity category, but different E_{CN}. The Twitter account with the largest E_{CN} is regarded as more similar with respect to the examined account. Therefore, another factor taken into consideration is the number of the intersected (common) entities E_{CN}. The resulting Common Coefficient (CC) of that entity category is calculated as presented in Equation 5.4:

$$E_{CC} = E_{Cf} * E_{CN}. \qquad (5.4)$$

By combining the two aforementioned factors into one equation, we calculate the Common Weighted Coefficient (CWC) of an entity category (Equation 5.5):

$$E_{CWC} = E_{Cf} * E_{CN} * E_{W}. \qquad (5.5)$$

The third factor that should be considered before calculating the Similarity Metric is the number of entity categories that the compared account has at least one entity in common with the examined account (label). This factor is used in order to adjust the metric by considering the existence of the number of the four distinct coefficients. Finally, the Similarity Metric (SM) is calculated by incorporating the four coefficients and the three factors into Equation 5.6:

$$\text{Similarity Metric} = (\text{hashtag}_{CWC} + \text{mention}_{CWC} + \text{URL}_{CWC} + \text{domain}_{CWC}) * \frac{\text{label}}{4}$$
$$\text{where: label} = \{0, 1, 2, 3, 4\},\ 0 \leq \text{SM},\ \text{SM} \in \mathcal{R}. \qquad (5.6)$$

All of the aforementioned coefficients and factors are based on the individual characteristics of each Twitter account, thus forming a dynamic and unique Similarity Metric for each pair of examined - compared Twitter accounts.

5.4.1. Case study

As a case study scenario we applied our proposed methodology on the Twitter account of the ex-minister and current member of the Greek parliament @adonisgeorgiadi. We selected this account since it is well-known, highly influential and active. We explicitly claim that we use this account for research purposes and we are not against or in favor with respect to its disseminated content. The aim of this case study is to discover its top-k similar accounts where *k*=15.

The resulting dataset is a graph which was queried in order to apply the proposed methodology in RDF format and it is publicly available[22]. The data can be queried through the provided endpoint[10] of the InfluenceTracker.com service under the named graph *http://influenceTracker/twitterGraph/full*. The information with respect to the case study was collected between Oct 13 and Oct 22 of 2014. A quick overview of the contents can be found in Table 5.1. The graph contains 90,578 Twitter accounts. All the information described in our ontology has been modeled for 2,423 of them. These were randomly selected from the mentions found in the captured tweets. It should be noticed that InfluenceTracker.com is an active site, therefore the reader of this document may find also additional accounts (new accounts are being continuously added). The remaining 88,155 accounts are followers or are being followed by the fully modeled 2,423 accounts. In addition, there are also 188,542 shortened URLs, while 72,931 among them have been transformed from tiny to typical URLs in order to retrieve their domains. For this operation we used the http://unshorten.it/ service. The transformed URLs are hosted by 8,402 unique domains. Finally, 38,020 hashtags and 59,160 images are modeled as these were contained in the captured tweets. All the presented data are modeled in nearly 2 million triples.

Table 5.1: The contents of the queried graph

Accounts	90,578	**URLs**	188,542	**Hashtags**	38,020
Full Accounts	2,423	**Full URLs**	72,931	**Images**	59,160
Simple Accounts	88,155	**Domains**	8,402	**Triples**	1,982,367

5.4.2. Case study results (depth=1)

The top-15 similar accounts of @adonisgeorgiadi according to our methodology are illustrated in Table 5.1. The nodes correspond to Twitter accounts, while the curving edges indicate a clockwise direction from the source node (@adonisgeorgiadi) to the target node. The thicker the edges the more similar we consider the connected nodes. The edges have the same color as their destination node. The presented network -as well as the others below- is created with the open graph visualization tool called Gephi (layout type: Yifan Hu).

[22] https://www.dropbox.com/s/i1ow3jt2dgdxzhn/itGraphFull.rar?dl=0

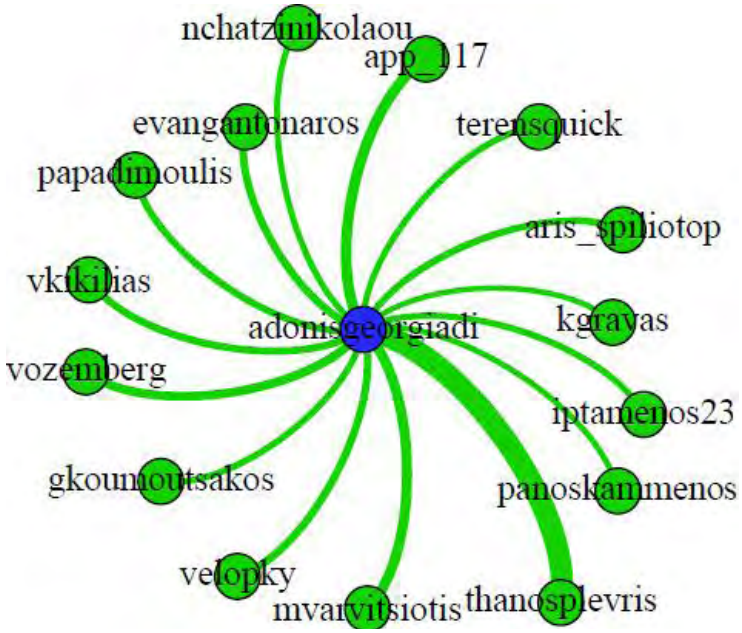

Figure 5.1: Case study similarity network (depth=1) - Thicker edges denote more similar Twitter accounts

We noticed that 12 out of the top-15 connected accounts, belong to current members of the Greek parliament, while the remaining three accounts belong to a well-known political journalist in Greece (@nchatzinikolaou) and to two persons (@app_117, @iptamenos23) who are posting tweets about the political situation in Greece and retweet messages of many politicians.

The first column of Table 5.2 presents these top-15 similar accounts, while the rest of the columns highlight the respective factors and metrics defined and presented previously. As it can be clearly seen, the presence of each distinct metric (and its respective value) affects the final Similarity Metric (SM) that is depicted in the final column of Table 5.2. For example, the account @evangantonaros is ranked as the third highest account according to CWC. Nevertheless, due to having one category less in common with the examined account (no common URLs are found), the "label" parameter of Equation 5.2 equals to 3, thus reducing the Similarity Metric by 33.3%. Finally, the @evangantonaros account is ranked as the sixth more similar account among the top-15.

5.4.3. Case study findings (depth=2)

As an extension of the previous experimentation, our next step was to increase the value of *depth* in order to discover the top-15 similar accounts of those displayed in Table 5.1. The proposed methodology was implemented iteratively for each one of the previous accounts. The produced network is illustrated in Figure 5.2.

Moreover, the resulting network consists of 107 Twitter accounts. Approximately two thirds of them belong to current members of the Greek parliament, as well as to persons who are actively engaged with political parties in Greece or even official political party accounts. The remaining accounts belong to journalists and to persons posting Tweets about the political situation in Greece.

These 107 nodes representing Twitter accounts are interconnected through 240 directed and weighted edges. Many of the accounts contained in the top-15 similar results are repeated, and therefore the actual number of appearing accounts is less than the maximum that can be achieved. In the presented scenario, where depth=2, the maximum number of unique nodes is 241 (that is 16 sets of top-15 similar accounts plus the root account @adonisgeorgiadi). It is worth noticing that the whole network is built on the 44.6% (107) of the maximum possible nodes (241), since the rest 134 appear again among the top-15 similar accounts one depth further.

An example of an account being in the other top-15 results is the examined-root account @adonisgeorgiadi. It appeared in the top-15 results of 13 similar accounts, thus its $\text{In-Degree}_{\text{top-15}}$ is 13. In a sense, this mutual similarity defines that the connected nodes are highly likely to be similar with at least an 86.67% probability value.

Table 5.2: The top-15 similar accounts of @adonisgeorgiadi along with their respective similarity metrics for the selected case study

Account	Hashtags		Mentions		URLs		Domains		CWC	Categories (out of 4)	SM
	total	common	total	common	total	common	total	common			
@thanosplevris	3	2	98	50	130	9	5	5	9.848	4	9.848
@app_117	18	5	59	29	52	0	23	18	6.303	3	4.727
@mvarvitsiotis	37	6	83	32	303	1	1	1	4.518	4	4.518
@vozemberg	40	3	65	26	93	3	2	2	3.913	4	3.913
@velopky	25	4	373	57	410	4	5	4	3.411	4	3.411
@evangantonaros	2	2	29	18	63	0	23	11	4.521	3	3.390
@vkikilias	59	4	74	29	309	0	3	3	4.292	3	3.219
@papadimoulis	55	5	128	38	402	0	2	1	4.067	3	3.050
@aris_spiliotop	20	3	37	19	65	0	21	12	4.052	3	3.039
@gkoumoutsakos	9	1	46	22	51	0	2	1	3.776	3	2.833
@iptamenos23	10	2	71	27	45	0	1	1	3.746	3	2.810
@terensquick	34	2	313	57	923	0	4	1	3.706	3	2.780
@nchatzinikolaou	68	7	309	47	1175	6	5	3	2.743	4	2.743
@panoskammenos	0	0	134	35	249	3	9	4	3.404	3	2.553
@kgravas	18	1	16	12	88	0	2	1	3.234	3	2.425

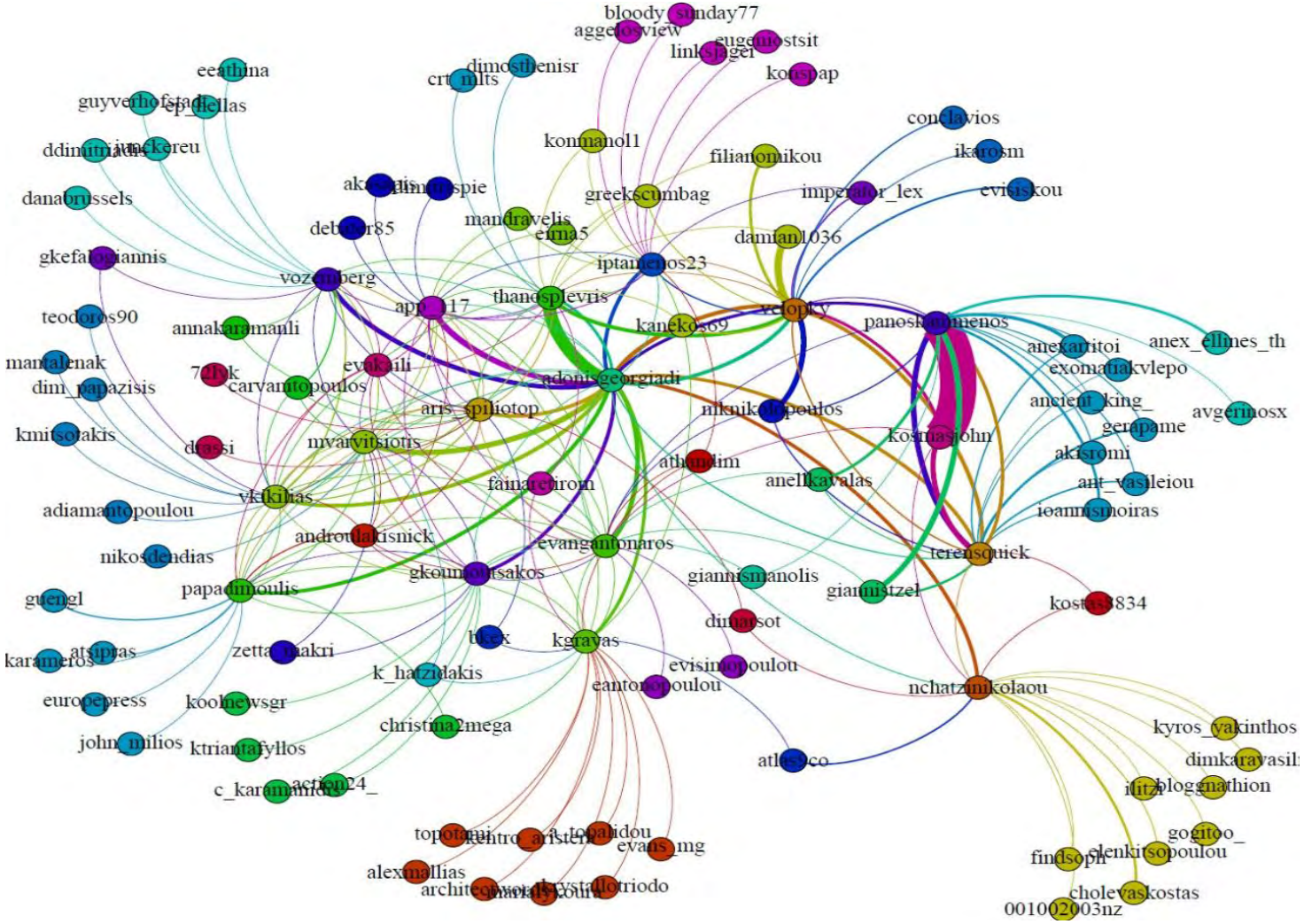

Figure 5.2: Case study similarity network (depth=2) - Thicker edges denote more similar Twitter accounts (root @adonisgeorgiadi)

5.4.4. Additional case studies findings

In this section, the results of two more case studies are briefly presented. Our framework has been also applied on the Twitter account of the current president of the European Commission @junckereu and the widely known news station @cnn. The aim is to discover their top-k similar accounts where *k*=15 and *depth*=2. The produced networks are illustrated in Figures Figure 5.3 and Figure 5.4 respectively.

The resulting network of @junckereu consists of 80 Twitter accounts. The vast majority of them are directly related and actively engaged to the European Commission, i.e. members of European parliament (MEP), related Cabinet members, political parties, ministers and politicians at the national level and journalists in the domain of politics. Table 5.3 presents the top-10 similar accounts of @junckereu along with a short description regarding them.

The resulting network of @cnn consists of 43 Twitter accounts. Approximately 85% of them are either accounts of TV shows or of working staff (correspondents, analysts, presenters and producers) of that news agency. Table 5.4 presents the top-10 similar accounts of @cnn along with a short description regarding them (Table 5.5).

Table 5.3: The top-10 similar accounts of @junckereu

Similar Account	Similarity Metric	Description
@eu_commission	8.05723	European Commission
@evp_at	5.520655	Political Party
@avramopoulos	5.4801927	MEP
@eucopresident	5.1954527	President of European Commission
@martinselmayr	4.9672847	Cabinet member
@martinschulz	4.022644	President of European Parliament
@anneschmtz	3.8992517	Retweets EU related tweets
@csv_news	3.7792974	Political Party
@evp_de	3.6083882	Political Party
@vozemberg	2.5673862	MEP

Table 5.4: The top-10 similar accounts of @cnn

Similar Account	Similarity Metric	Description
@cnnbrk	164.6085	Service of CNN
@cnnvideo	65.3353	Service of CNN
@earlystart	64.01588	CNN TV Show
@crossfire	61.31107	CNN TV Show
@cnntonight	39.4743	CNN TV Show
@cnncameraman	39.18925	Works at CNN
@christinacnn	37.17193	Works at CNN
@cnn_stevealmasy	34.63642	Works at CNN
@cnnsitroom	30.94307	CNN TV Show
@jaymcmichaelcnn	30.30539	Works at CNN

Table 5.5: The maximum number of accounts and the unique ones inserted into the network

Depth	Maximum Accounts	Increase
0	1	1
1	16	15
2	241	91
3	3841	258

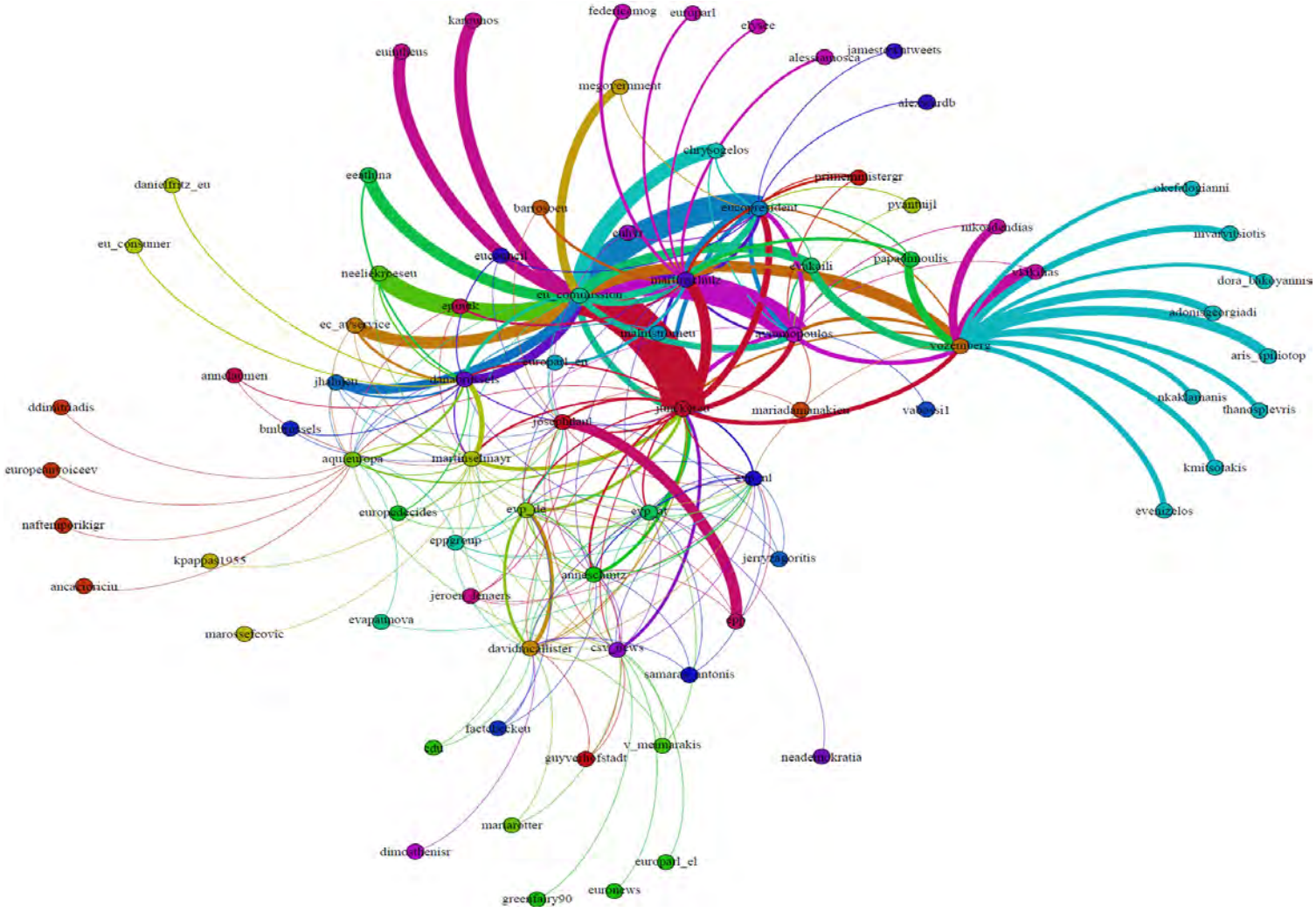

Figure 5.3: Case study similarity network (depth=2) - Thicker edges denote more similar Twitter accounts (root @junckereu)

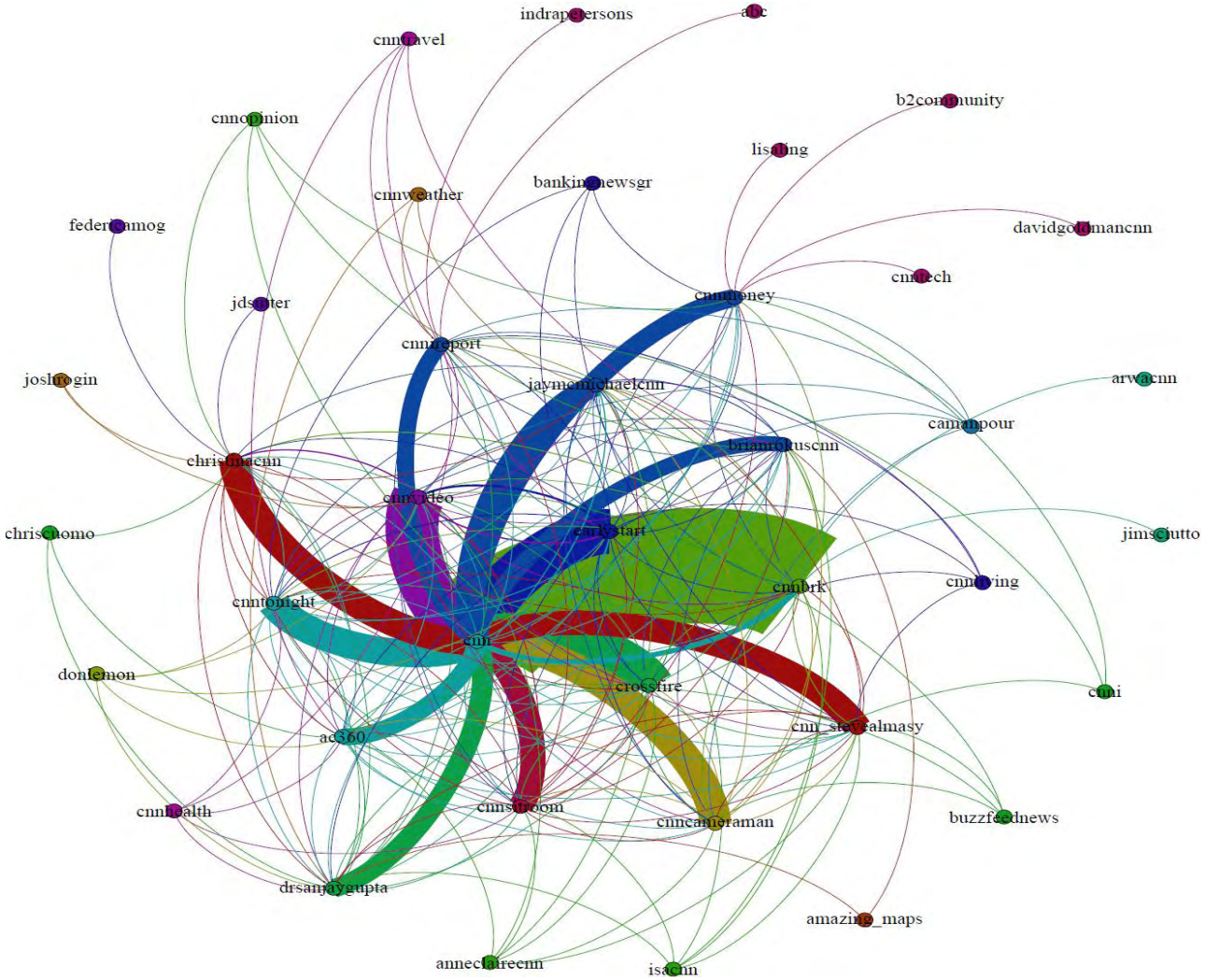

Figure 5.4: Case study similarity network (depth=2) - Thicker edges denote more similar Twitter accounts (root @cnn)

5.5. Labeling Twitter Accounts

There are 44 thematic categories in total used by the proposed algorithm for labeling the Twitter accounts. These can be viewed in Table 5.9 (in Appendix B: DBpedia Categories). 40 of them have been derived from DBpedia while the remaining four are proposed by us. From those 40 categories, the 36 have been inspired by the values of the "rdf:type" property of the related DBpedia resource, while the remaining four from the values of the "dbp:position" property. The first property is used to state that a resource is an instance of a class, while the latter to state the political position of a resource. This property was used in cases where we wanted to introduce a thematic category but no suitable DBpedia ontological classes could be found.

The rationale behind the selection of these thematic categories is based on two factors. The first is the classes, these are the values of the "rdf:type" property, which can be grouped under the same category, and the second is the number of occurrences in our dataset.

It is common that different ontological schemes introduce their own classes for the semantic representation of the same or of very similar domain. For example the classes dbo:Company, yago:Company108058098, and schema:Organization refer to companies. Moreover, it is possible to find in a single ontology classes representing broader or narrower concepts of the same domain. For example, the classes dbo:SportsLeague, dbo:SoccerLeague, and dbo:AmericanFootballLeague, despite not representing exactly same concept, they are all referring to the broad domain of "League".

In order to discover the most dominant thematic categories we used a SPARQL query in order to display all the classes contained in the dataset and counted their occurrences. Continuing, we excluded the generic classes (e.g. owl:Thing, dbo:Agent, foaf:Person, yago:Whole100003553) and grouped the remaining classes according to their domains as described in the previous paragraph. Finally, we aggregated the occurrences of each class belonging to a domain as well as the number of classes representing that domain, and we calculated the average occurrence of each class per domain. The domains having average greater than one were used as the thematic categories for labeling the Twitter accounts. These can be viewed in Table 4.3 (in Appendix B: DBpedia Categories). On average each thematic category is based on 19.5 classes having an aggregated number of occurrences equal to 46.67, which is equivalent to 2.56 occurrences per class.

5.5.1. A Thematic Category Labeling Algorithm

In this section, we present the proposed methodology for labeling the Twitter accounts with the thematic categories presented in Section 5.5, in order to not only classify them into communities, but also to enrich the social data with valuable information from DBpedia. The "Thematic Category Labeling Algorithm" demonstrates the benefits of the semantic technologies and the LOD cloud, thus increasing the value of the available data by providing answers to sophisticated queries (e.g. return the top-10 members of political parties according to the number of their followers). Moreover we analytically display and discuss the results of the proposed algorithm.

As already mentioned, this work extends the work described in Section 5.4, which aimed at discovering similar accounts in Twitter, in terms of their disseminated entities (mentions, replies, hashtags, photos, and URLs). One of the outputs of this process was files presenting analytically for each account its similar ones along with their Similarity Metric (SM) value. The data in these files are used for creating weighted graphs of similar accounts.

The "Thematic Category Labeling Algorithm" (TCLA) is applied to all of the accounts found in our system by utilizing their *top-k* similar ones. It consists of three phases, as presented below.

Phase 1: Thematic Category Labeling

1.1. Initially the accounts with a relation to a DBpedia resource describing them are labeled with one or more thematic categories, depending on the DBpedia data.

1.2. An output file of an account derived from the Similarity Recommendation process, as described in Section 5.4, e.g. AdonisGeorgiadi.txt, is used as input. This is the *root* account.

1.3. From that file, we acquire the *top-k* similar accounts of the root and their SM value (e.g. ThanosPlevris - 42.04). These are the accounts-*sources* which will inherit their thematic categories along with their respective weights (Equation 5.7) to the root (see step 1.3.b). At the end of this step the Initial Labeling will be completed and the following information will be available:

$$\text{weight} = \frac{\text{SM}_{\text{similar}}}{\text{SM}_{\text{root}}} \qquad (5.7)$$

1.3.a. The "sources" of the root account (e.g. AdonisGeorgiadi (self SM: 2707.9): {[ThanosPlevris, 42.04], [kMitsotakis, 21.29]}).

1.3.b. The thematic categories of the root along with their "Thematic Category Weighted Score" (TCWS) (Equation 5.8) (e.g. AdonisGeorgiadi: {[Politics, 40.21], [Right Wing, 25.29], [Entertainment, 0.32]}). This score is affected by the value of the SM between the root and the source account. For avoiding outlier values in case that the TCWS is greater than 10, it is adjusted as "$\log_{10}$(TCWS) * 10". The self TCWS of an account is always equal to 1.

$$\text{TCWS} = \begin{cases} \text{weight}, & i = 0 \\ \text{TCWS}_{\,i-1} + \text{TCWS}_{\,i-1} * \text{weight}, & 1 \leq i \leq 5 \end{cases} \qquad (5.8)$$

1.4. The results of the Steps 1.2 and 1.3 are stored temporarily.

1.5. Steps 1.2 to 1.4 are repeated until all the output log files have been used as input.

Phase 2: Iterative Thematic Category Labeling

During this phase, the basic notions of Step 1.3 are iteratively executed for all of the accounts for a predefined number of iterations. In more detail, this phase includes the following steps.

2.1. An account is selected in order to be iteratively labeled. This is the *root* account.

2.2. The input of this step for each "root" account are: i) the results of Phase 1, namely its thematic categories along with their TCWS, and ii) its *top-k* similar accounts along with their SM value. The basic notions of Step 1.3 are executed. At the end of this step the previous results are enriched and adjusted, while new labels can be added and the TCWS values of existing ones can be increased.

2.3. The results of this i^{th} iteration are stored.

2.4. Steps 2.1 to 2.3 are repeated until all the accounts are iteratively labeled.

2.5. Steps 2.1 to 2.4 are repeated *i* times.

Phase 3: Thematic Categorization Results

3.1. As input, the results of Phase 2 for all the accounts are acquired.

3.2. An account is selected. The results of the iterations of Phase 2 are used in order for the account's final thematic categories to be discovered. This is achieved by applying the following:

3.2.a. The thematic category label having the greatest TCWS, called $Dominant_{TCWS}$, is found. In the example presented in 1.3.b. the $Dominant_{TCWS}$ label is "Politics".

3.2.b. The thematic categories having TCWS less than 20% of the $Dominant_{TCWS}$ are ignored as outliers.

3.2.c. If the $Dominant_{TCWS}$ label of an iteration is two times or more than the respective of the previous one, then the labeling process ends and the account is labeled with these thematic categories. In the example presented in 1.3.b the final thematic category labels are: Politics (40.21) and Right Wing (25.29).

3.3. Step 3.2 is repeated until all the accounts are labeled with their final thematic categories.

3.4. The final thematic category labels of the accounts are displayed and stored.

5.5.2. Results

As already mentioned, the proposed algorithm is applied iteratively to the *top-k* similar accounts of each one in our system for *i* iterations. In our case, *k* has been set equal to 50 and *i* equal to 5.

During the period we applied our methodology 986 Twitter accounts were registered in the InfluenceTracker.com system. 408 are related to 406 DBpedia resources providing more information about them. Those were labeled with one or more thematic categories according to the data in their resource.

As seen in Table 5.6, these 408 initial accounts were labeled in total with 777 thematic categories, leading on average to 1.9 thematic categories per account. For 19 accounts no similar ones were found, due to the lack of entities (mentions, replies, hashtags, photos, and URLs) contained in their tweets. After the application of TCLA, 962 accounts were labeled with 4,099 thematic categories in total, leading on average to 4.26 labels per account. The average variation between the initial and the final thematic categories per account is 2.36 labels leading to an increase of 223.74%.

In order to further investigate the results, we compared the labeling of the thematic categories between the group of the initial accounts and the newly-labeled ones after the application of TCLA (Table 5.7). The first group consists of 408 accounts being labeled with 1,856 thematic categories in total, 4.55 tags per account on average, while the latter consists of 554 accounts being labeled with 2,242 thematic categories in total, 4.05 tags per account on average. As can be seen in Table 5.8, the average variation of the group of the initial accounts before and after the application of TCLA is 2.65 thematic categories, leading to an increase of 238.87%.

Table 5.6: Metrics before and after the TCLA

Stage	Thematic Categories	Accounts	Average
Start	777	408	1.9
End	4,099	962	4.26
Variation	**3,322**	**554**	**2.36 (+223.74%)**

Table 5.7: Final Labeling of group of Initial and Newly-Labeled Accounts

Group	Thematic Categories	Accounts	Average
Initial	1,856	408	4.55
Newly-labeled	2,242	554	4.05

Table 5.8: Metrics of Initial Accounts Before and After the TCLA

Stage	Thematic Categories	Average
Start	777	1.9
End	1,856	4.55
Variation	**1,079**	**2.65 (+238.87%)**

108 displays a tag cloud[23] which contains the top-12 thematic categories of the account @AdonisGeorgiadi. In the tag cloud the most dominant labels are represented with bigger fonts. Half of the labels are related to the domain of politics. This account before the application of TCLA was labeled with "Politics" and "Right Wing". After the application of the proposed algorithm the account was labeled with more thematic categories which were derived from its similar ones.

Figure 5.5: A tag cloud containing the top-12 thematic categories of an account

5.6. Similarity Recommendation Evaluation

The purpose of this section is two-fold. We first want to evaluate the results of the case study described in Section 5.4.1. Then, in order to further evaluate the Similarity Metric, we describe a generic evaluation, which involves subjective user ratings for the results obtained from the proposed metric.

5.6.1. Case study evaluation

The experiment has shown that on average the In-Degree$_{top\text{-}15}$ of a "root" account is almost equal to 12. That suggests that there is an 80% probability (12 out of 15) of the "inverse" similarity relation to exist between the examined account and its top-15 similar ones. Simply put, if an account B is in the top-15 similar ones of the examined A then there is an 80% probability of A being in the top-15 similar accounts of B. This fact reflects the dynamic nature of our Similarity Metric, since it is based on the individual characteristics of each account, it is almost unique for each pair of examined - compared Twitter accounts.

As already mentioned, when expanding the depth of the network the theoretical maximum number of nodes is not reached, mainly due to mutual similarities among the nodes. Specifically, as the values of *k* and *depth* increase, the number of the unique nodes in the network rapidly decreases, while the total number of nodes increases at a lower rate.

[23] Created with: https://www.jasondavies.com/wordcloud/

Figure 5.6 presents the theoretical maximum number of accounts (per depth) versus the actual unique ones inserted into the network. The horizontal axis represents the depth with respect to the initial node (root), while the vertical axis represents the number of examined accounts. The diagram depicts two dotted lines, along with their trend lines of exponential type. The blue dotted line represents the theoretical maximum number of accounts that needs to be explored, while the red dotted one the number of unique accounts that were eventually explored into the network. Table 5.5 presents all the respective values according to the depth with respect to the root node.

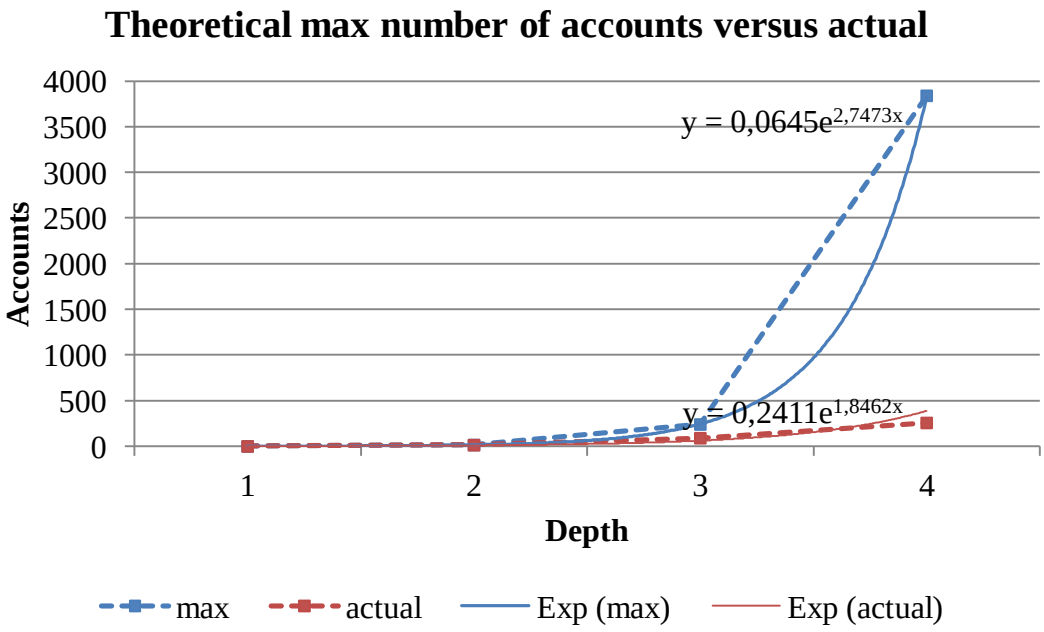

Figure 5.6: The theoretical maximum number of accounts (per depth) (blue-colored curves) versus the actual unique ones (brown-colored curves) inserted into the network - Trending behavior is according to exponential type (values are depicted in Table 5.5)

Finally, we also noticed that the more top-k similar accounts are examined the less unique accounts are discovered, and the more closed walks (cycles) are found in the generated network. In this case study, a cycle[24] of length 5 that starts and ends from the @adonisgeorgiadi node is {@adonisgeorgiadi → @thanosplevris → @vozemberg → @vkikilias → @aris_spiliotop → @adonisgeorgiadi}. Such a cycle reveals a community of similar accounts (all accounts are politicians belonging to the same or adjacent political parties).

It is worth noting that such a kind of communities follow a power law distribution. Figure 5.7 presents the cycle (community of similar users) distribution of a network that consists of 365 unique accounts, which were discovered after the methodology was applied for depth=3 and k=15 (sum of last column of Table 5.5). The vertical axis represents the number of closed walks, while the horizontal axis represents their size. Among 365 accounts, a total of 531 cycles were revealed. Figure 5.7 depicts the respective power law distribution.

[24] A cycle is a unique closed walk (across different nodes) that starts and ends from a distinct node.

Two of these cycles have length equal to 68, which is approximately 19% of the total nodes. As the number of the top-k accounts increases, the average number of nodes per cycle also increases. Figure 5.8 depicts the average nodes per cycle distribution of the resulting similarity networks after each depth. The vertical axis represents the amount of the average nodes per cycle, while the horizontal axis the depth. In the presented case the average weighted cycle size is approximately 24.

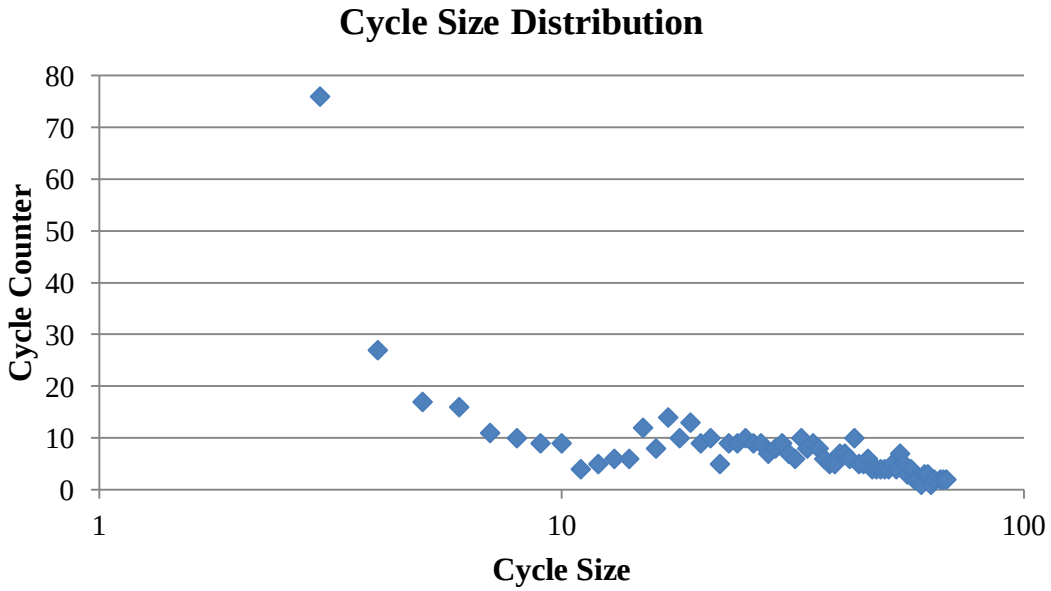

Figure 5.7: The closed cycle size distribution of a network (356 unique accounts - 531 closed cycles)

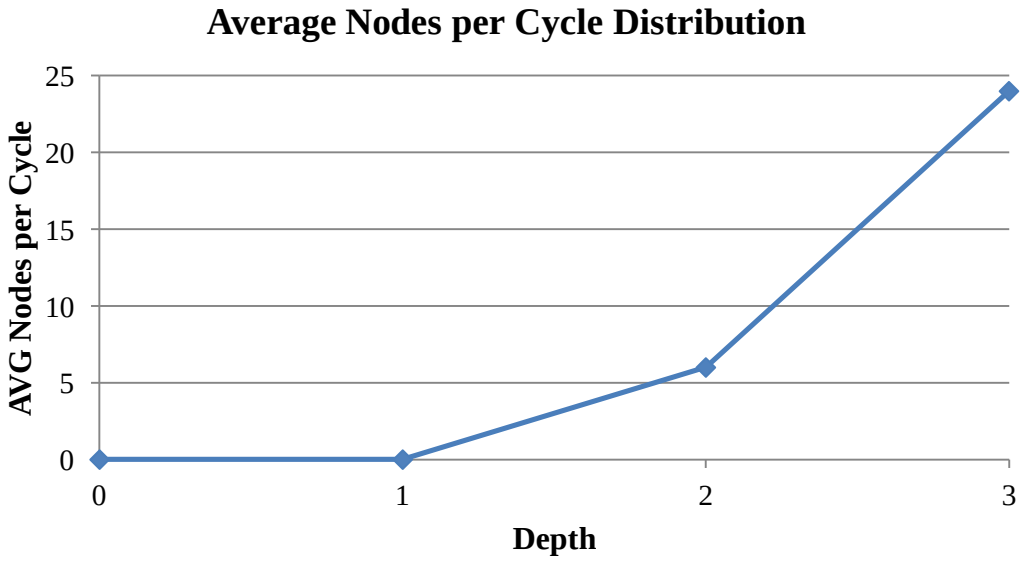

Figure 5.8: The average nodes per cycle distribution for each depth

5.6.2. Evaluation against user ratings

One of the functionalities offered by Twitter to its users is the recommendation of other accounts (as similar) to be followed[21]. These suggestions are personally provided to the users and are mainly based on the users' contacts, e-mails, locations, followers and followees, as well as on other public profile information. Very little attention has been given to the content itself (e.g. text or Twitter entities). Moreover, these suggestions are only visible to the account owners and cannot be retrieved using the Twitter API. As a result, it was obvious that we could not evaluate our methodology having as ground truth the respective recommendations provided by Twitter. Thus, in order to further evaluate our methodology, we describe here a generic evaluation over subjective user ratings.

Moreover, for the purpose of this evaluation, 22 postgraduate students from an MSc course class at the University of Thessaly were engaged. Their task was to subjectively rate the similarity results provided by our methodology. Each student was asked to select an initial root node and then evaluate the similarity network derived when seeking the top-5 similar accounts when the depth search equals to 3. Each individual had to explicitly rate how similar two accounts are -for all the various cases in the resulting network- under a five-point Likert scale, as indicated below:

1. Strongly disagree (totally unsimilar accounts)
2. Disagree (rather not similar accounts)
3. Neither agree nor disagree (I cannot judge - neutral)
4. Agree (the accounts tend to be similar)
5. Strongly agree (I am sure. These accounts are similar)

22 distinct case studies were evaluated by each individual, in the same sense as the case study described in Sections 5.4.1 and 5.4.2. However, we set the search depth equal to 5, while in order to keep the amount of possible ratings between nodes in manageable levels, we reduced the top-k examined accounts setting k equal to 5.

In Figure 5.9, we can see the points that indicate the average ratings of the evaluators between nodes and according to their distance in the resulting similarity network. As distance, we define the number of hops between the compared nodes. We noticed that when the distance between compared nodes increases, the average subjective similarity rate value (in the five-point Likert scale) decreases. This result was somehow expected since in a resulting network of the top-k similar accounts of the top-k similar accounts and so forth (according to the selected search depth), the higher similarity values between nodes tend to appear in nodes with lower distances. We also noticed that for low distance values (up to 2) the mean ratings are above 4, denoting that our Similarity Metric works efficiently enough according to the evaluators' opinion.

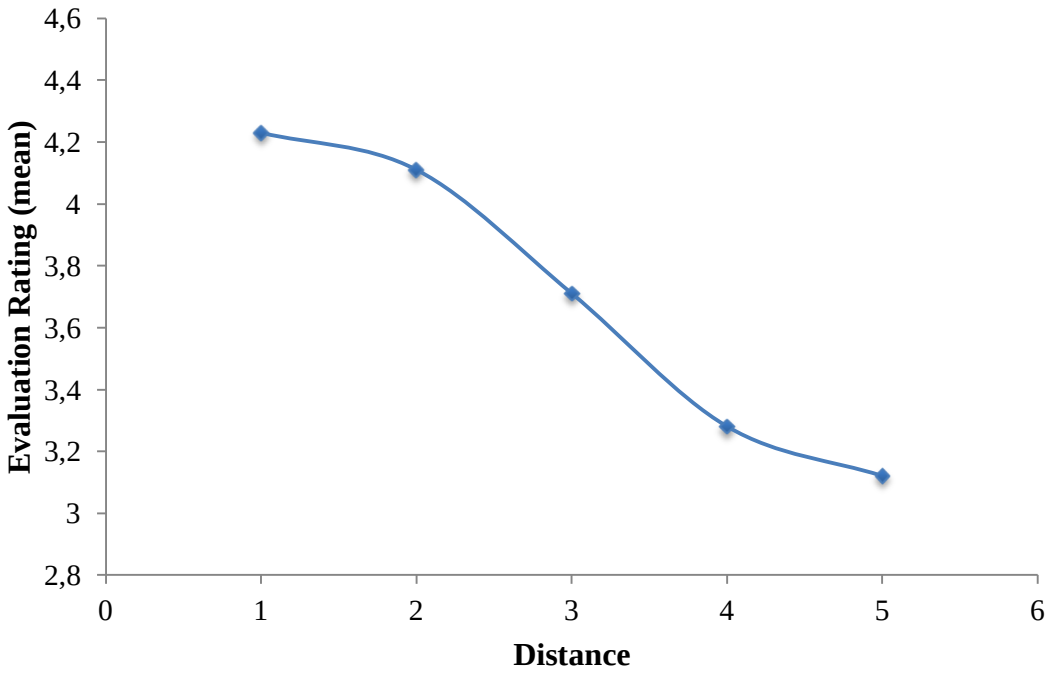

Figure 5.9: Mean rates (from the evaluators) versus Distance in Similarity Network

Chapter 6. Semantic social search

6.1. Introduction

Microblogging - a "light", rather live, version of blogging - is considered to be one of the most recent social raising issues on the Internet, being one of the key concepts that brought the Social Web to the broad public. The main characteristic of microblogging is the fact that posts are produced almost in real-time and are strictly limited to a specific and rather small number of characters, such as short sentences, term concatenation or shortened URLs that point to hyperlinks with web and multimedia content. Microbloging comprises many very brief updates that are presented to the microblog's readers in reverse-chronological order. Motivated by its increasing popularity, among many microblogging services we focus on the Twitter social network, where microblogs are known as tweets.

In web information retrieval, the effectiveness of search engines strongly depends on whether users can express their information needs through the terms they submit. However, submitting the right queries is not an easy task since queries are usually short, not written in natural language, and -mostly- their terms are ambiguous. Many proposed methods offer meaningful query suggestions, usually by employing knowledge extraction methods from browsing history records or search logs. However, very few consider time as an important parameter related to the actual meaning of a query term. In this chapter, we do not tackle query suggestion in the traditional way, but we provide time-aware suggestions according to the most viral terms that appear in Twitter along with the user's query.

The main contribution of this chapter is the effective suggestion of microblogging social content (called hereafter as *Twitter Entity/Entities - TE/TEs*) that manage to become viral in time, given a user query; the more viral the social content is, the more relevant the suggestions are. Our ultimate goal is to provide users with a way to enter any type of query and retrieve accurate, relevant and popular (viral) Twitter Entities suggestions that would semantically "fit" to their information needs. In order to measure virality, we extend the capture-recapture methodology, which is mainly used for estimating population properties (e.g. birth/survival rates) in real-life biological experimentations. In our work, the concept of virality in a social content is considered the same as survivability in the animal populations under study. The concept of social content on the other hand is directly related to TEs (hashtags (#), user mentions (@) and URLs) and should not be related to named entities or Wikipedia concepts as considered in most papers in the related literature of text mining and information retrieval[25] (e.g., see the work described in (Spina et al., 2012)).

It is true that various research works on microblog posts analysis and extraction of meaningful information from them in a (semi-)automated manner have been considered recently in the literature. Nevertheless these approaches are quite different from our work. As the interested reader will see within next sections, related research on *query suggestion* (Mishne et al., 2013) is highly related to *query expansion* (Massoudi et al., 2011), *query substitution*, *query recommendation* or *query refinement* tasks. In this work, we deviate from the traditional query suggestion

[25] Thus, it should be clear that whenever we mention the term "entity" in the manuscript, we refer to Twitter Entity/Entities (TE/TEs), unless otherwise explicitly stated.

proposals in a sense that users have their queries expanded directly from Twittersphere, without having their queries or browsing history processed by search engines. In addition, another important difference against related query suggestion techniques, focused on web search and the real-time variance of the problem at hand, is the narrow time frame considered herein in which suggestions have a maximal impact. For this work we were also motivated by the facts that a) microblogging social content annotation is provided directly in real-time by users worldwide and b) the more this annotation becomes important or so-called "*viral*", the more semantically related it becomes with a recent trend, the top news, a thematic categorization, etc.

To summarize, this chapter provides the following contributions:

1. We present a query expansion methodology, for effectively suggesting timely viral microblogging social content (hashtags, mentions, URLs).
2. We measure virality, by extending the capture-recapture methodology, which is mainly used for estimating population properties in real-life biological experimentations.
3. We deviate from the traditional query suggestion proposals by expanding the users' queries directly from Twittersphere, without processing any past queries or browsing history.

The rest of this chapter is outlined as follows. In sections 6.2 and 6.3 we provide an overview of the literature within the query suggestion field, emphasizing on related works within the social sphere. Section 6.4 provides an overview of the methodology we adopt, as well as the basic steps of our proposed query suggestion method. In Section 6.5 -and in order to clearly show how our query suggestion expansion mechanism works- we describe the results of two real-life scenarios (case studies). In addition, we evaluate our results against four famous web news services (Google News, Yahoo! News, Bing News, and Reuters). Finally, in Section 6.6 we further evaluate our approach by subjective comparisons with respect to the Google Hot Trends service, as well as against a cluster labeling and a microblog retrieval task, and we provide comparative results.

6.2. Information search and retrieval in microblogs

In general, Twittersphere consists of the so-called tweets or microblogging posts (Efron, 2010), where this large amount of real-time tweets per day is highly attractive for information retrieval research. Within that social sphere, the context of query suggestions must be in real-time, i.e. results need to be temporally relevant and timely (Mishne et al., 2013). Microblogs form a rather special category of user-generated data: they typically contain two major characteristics that seriously affect the expressiveness of linguistic analysis techniques, namely: a) they contain strong vernacular (acronyms, spelling changes, etc.) and b) they do not include any memorable repetition of words. More specifically, (Massoudi et al., 2011) study a Twitter-based retrieval model by considering the model with textual quality and Twitter specific quality indicators. (Naveed et al., 2011) combine document length normalization in a retrieval model to resolve the short texts sparsity problem in the

case of tweets. Motivated by the observation that in a typical microblog users tend to retrieve meaningful information through a query formulation, researchers focus on each post's characteristic features (Huberman et al., 2009), whose quantitative evaluation could potentially affect the way in which the relevance between the user query and its returned results may be calculated.

The fact that microblog posts contain hashtags is also exploited in the literature in the direction of acquiring information that the user "is not aware of" and to formulate queries that the user "does not know how to express" (Biancalana et al., 2013). In a representative approach, (Efron, 2010), the researcher given a query attempts to statistically identify a number of hashtags relevant to the given query that may be used to expand it and lead to better results.

Furthermore, the observation that microblog posts are created during an actual event and contain comments and/or information directly related to it, leads to various event detection research efforts, based on posts and/or hashtags, as the one in (Packer et al., 2012). Moreover, the authors in (Poghosyan and Ifrim, 2016) presented a story-tracking framework of modeled as a pattern mining and real-time retrieval problem. The most popular news stories, assigned with hashtags, are detected by mining frequent hashtag pattern sets. Using query expansion on the original hashtags new story articles are retrieved. The pattern set structure enables the hierarchical and multiple-linkage representation of the articles.

Last but not least, query expansion techniques using the users' interest profiles have been proposed. Such a study is presented in (Reda et al., 2011), which takes into consideration the similarity between tags composing a query and the social proximity between the query and the user's profile. Its aim is to assist users by refining and formulating their queries and by providing them with information relevant to their interests. The study in (Zhou et al., 2012) describes a query expansion framework that takes into account the users' preferences which are derived by analyzing microblog posts and hashtags related to the targeted users. Finally, the authors of (Celik et al., 2011) propose a framework for enriching Twitter messages with semantic relationships by analyzing Twitter posts. These relationships are identified among persons, products, and events and are utilized in order to provide query suggestions to the users.

6.3. Query manipulation works

Typical microblog query manipulation research problems include both query analysis and expansion and query suggestion approaches. There are still some distinctive differences between these two techniques. A query expansion task is typically used transparently from the end-user and internally within a search engine mechanism, whereas a query suggestion is exposed to its end-users and therefore can use additional explicit information to its aid. In this manner, the authors in (Bandyopadhyay et al., 2012) attempt to improve weak ad-hoc queries through a process they call "web assistance", by exploring standard query expansion approaches and by utilizing external corpora as a source for the query expansion terms, namely pages derived from the Web and their titles. The study in (Efron, 2010) showed that for a Twitter microblog collection, hashtags may be predicted using query expansion techniques, by restricting the added query terms to those candidates that are hashtags, stripping candidates of their leading "#" character. In another more recent approach

(Kumar and Carterette, 2013), the authors take into account the fact that most of the existing models for Information Retrieval do not take the very important time aspect into account and focus on Twitter search models; they utilize time-based feedback and a simple query expansion by using highly frequent terms in top tweets as their expanded terms. In (Massoudi et al., 2011), the authors propose an efficient dynamic query expansion model for microblog post retrieval, utilizing a language modeling approach to search microblog posts by incorporating query expansion and certain "quality indicators" during the matching process. The latter is very interesting since several typical microblog characteristics may be exploited as quality indicators, such as temporal (Lee and et al., 2010) or topological ones.

In the case of actual query suggestion tasks though, the problem at hand becomes slightly different and its complexity increases as all the current major web search engines and most proposed methods that suggest queries rely solely on search engine query logs to determine their possible query suggestions. Although there are some research works on the topic in general, the consideration of the very important temporal parameter is rarely tackled, due to the fact that it is considered much more difficult to effectively suggest relevant queries to a recent search query, which has absolutely none or very few historical evidences in the aforementioned type of query logs. In this manner, (Li et al., 2013a) introduce the notion of fresh queries, trying to offer an effective query suggestion methodology for fresh search queries. Nevertheless they utilize word frequency statistics to extract a set of ordered candidate words for suggestions and not the most common Twitter entities (namely: hashtags, mentions and URLs) appeared in tweets, as we also propose in our work. Moreover, other attempts, like (Hu et al., 2011), implement empirical evaluations on a selected Twitter dataset in comparison to crawled hot queries published by Google Trends[26] for a given period of time, in a manner similar to the herein proposed approach. Finally, study (Mishne et al., 2013) presents the architecture behind Twitter's real-time related query suggestion and spelling correction service, as a case study illustrating the challenges of real-time data processing in the era of "big data" and argues that query expansion terms may be considered to be explicitly controlled by the user in an early form of query suggestion.

6.4. A Query Suggestion mechanism

Having discussed most of the related research works in the field, in this section we present the basic aspects of our proposed methodology. The microblogging service used in this work is Twitter. Thus, herein discussed Query Suggestion is based on the most common Twitter Entities (TEs), namely hashtags (#), mentions (@), as well as the links appearing in tweets. In order to keep the restrictions of 140[27] characters per tweet, Twitter uses a specific service that shortens the hyperlinks to 22 characters (tiny URLs).

Our proposed Query Suggestion mechanism has two steps. At first, we measure the virality (or survivability in our model) of the suggested TEs within a specific time and given a user's query, thus forming a cluster of candidate suggested terms for the Query Suggestion Set (Section 6.4.1). Then we calculate their ranking order among the suggested TEs (Section 6.4.2).

[26] https://trends.google.com/trends
[27] On September 2017 the limit was increased from 140 characters to 280.

6.4.1. Survivability factor - clustering suggested terms

Prior to describing our methodology, we need to introduce the framework of capture-recapture experiments, which are used mainly in wildlife biological studies (Pollock et al., 1990). In these experiments, animals, birds, fish or insects (subjects of investigation) are captured, marked and then released. If a marked individual is captured on a subsequent trapping occasion, then it is mentioned as a "recaptured" instance. The number of the marked and recaptured individuals can lead to an estimation of the total population size, as well as the birth, death and survival rates of each species under study. In our methodology, we specifically employ the Pollock's Robust Design model for clustering the candidate suggested terms (TEs in our case) that consist of the Query Suggestion Set. The Pollock's Robust Design model helps us calculate the survivability factor (also called survival probability - φ), which creates the cluster of the candidate suggested terms. In our paradigm, social content dynamics are considered analogous to the population dynamics. More specifically, a birth is the appearance of a new TE, while high survivability rates in these entities reflect high levels of virality.

To further elaborate on this aspect, the methodology of capture-recapture in real-life experimentations is briefly presented in the following: the sampling process is divided into k primary sampling periods, each of them consisting of l secondary sampling periods. At this point we have the distinction of the "open" and "close" models. In the first case, we assume that we can have births, deaths and/or migration incidents within the population under study, while in the latter the population and its evolution remains constant. In our model, we consider the "open" model among primary sampling periods and the "close" model among secondary sampling periods (Pollock et al., 1990). The basic measurements are conducted during a secondary sampling period, where a set of different individuals is trapped. Then these individuals are marked - keeping in parallel a history record of them- and then released back to their environment. After a specific time interval, the second secondary sampling period occurs and so forth until the end of the last l secondary sampling period. Secondary periods are near and quite short in time, while trapping occasions are instantaneous for assuming that the population under study is closed. However, longer time intervals between primary sampling periods are desirable so that evolution events can occur (e.g. survival, movement, and growth).

In our paradigm the trapping occasions corresponds to the query term (seed) we want to extend. Primary sampling periods consist of *16* secondary sampling occasions. In each of these *16* distinct samplings we capture and mark some entities with probability p. This probability value is the proportion of marked or total marked and unmarked Twitter entities that are captured during a sampling occasion, thus ensuring that all the secondary samplings are conducted in a "close" pool of instances and under the basic principle of the Pollock's model. Then, by investigating the recaptured instances, we calculate the survival probability φ of the examined entity according to Equation 6.1, where (M_i - m_i) defines the marked entities not captured during the i^{th} sampling period, while R_i is the number of entities captured at the i^{th} period, marked, and then released for possible recapture in future samplings. Moreover, M_i is the number of marked entities in the population at the time where the i^{th} sample is collected ($i \in \mathbb{Z}$, $1 \leq i \leq k$, $M_1=0$) and m_i stands as the number of the marked TEs captured in the same sample:

$$\varphi_i = \frac{M_{i+1}}{M_i - m_i + R_i} \qquad (6.1)$$

Figure 6.1 highlights the basic structure of the capture-recapture model we follow. The red boxes correspond to a primary sampling period, which is divided into 16 secondary sampling periods (blue boxes) and each period lasts for 1 minute. As mentioned above, the secondary sampling periods are quite important for our methodology since we can measure how viral a TE is. Now, in order to get a clear insight on how the algorithm works, let us see all these steps with an example. We assume that the user wants to have some suggestions next to the initial query term "Schumacher".

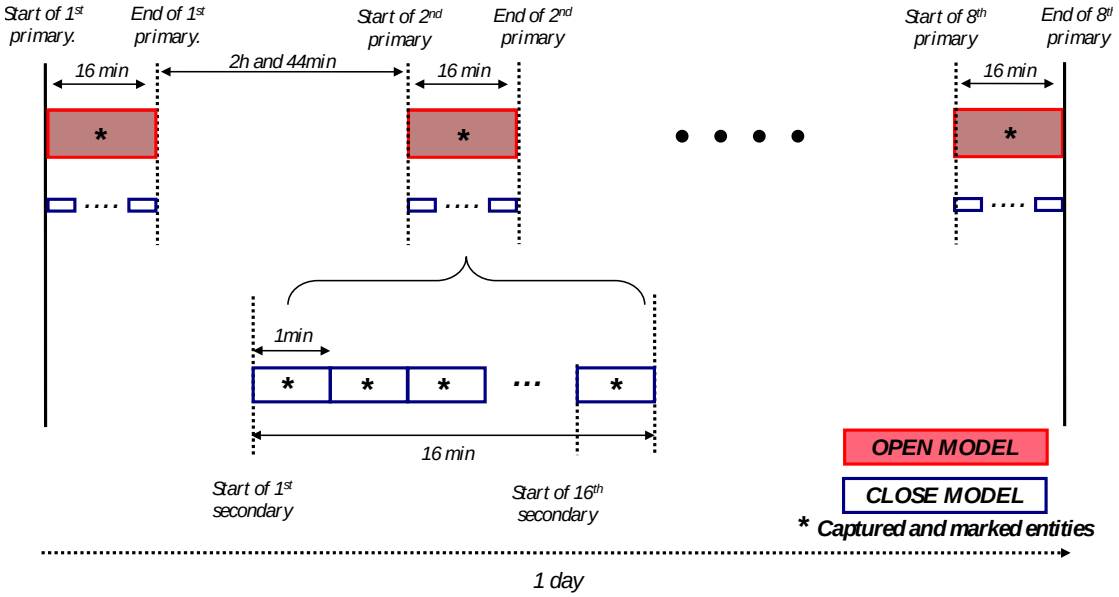

Figure 6.1: Structure of the conducted capture-recapture experiments (primary / secondary sampling periods)

As the flowchart of Figure 6.2 depicts, during every separate minute after the system receives the query "Schumacher" (also called seed term), we fetch through the Twitter API all the related tweets that contain this specific term. Then, from these tweets, we extract all the TEs (hashtags, mentions and URLs) and, finally, we select some of them according to the proportion of marked recaptured versus unmarked captured TEs during the subsequent sampling occasions (in this example the first value between the 1st and the 2nd secondary period was measured close to *p=6.25%).* As a result, the more a marked TE manages to appear again and again in the 16 subsequent samplings, the more viral is considered and becomes a strong candidate for a suggested term. On the contrary, the fewer times a TE appears within the short-time subsequent sampling occasions, the less viral and important is considered and it will eventually be ignored as a suggested term. Finally, upon completion of the first primary sampling period, we select the top-k% TEs (in this example *k*=10*)* that managed to appear in most of the 16 separate secondary sampling periods of the first primary sampling period. These entities reflect a significant trend behavior with respect to the seed term. Especially the hashtag #getwellsoonmichael presented the

higher survival rate reaching 94% frequency of appearances (appeared in 15 out of the total 16 samplings of the primary period). The rest top appearances were measured for #f1, #virus, #michael, #schumi and #legend, all related to the famous formula 1 driver who had a serious ski accident on December 29 of 2013 in the French Alpine resort of Meribel.

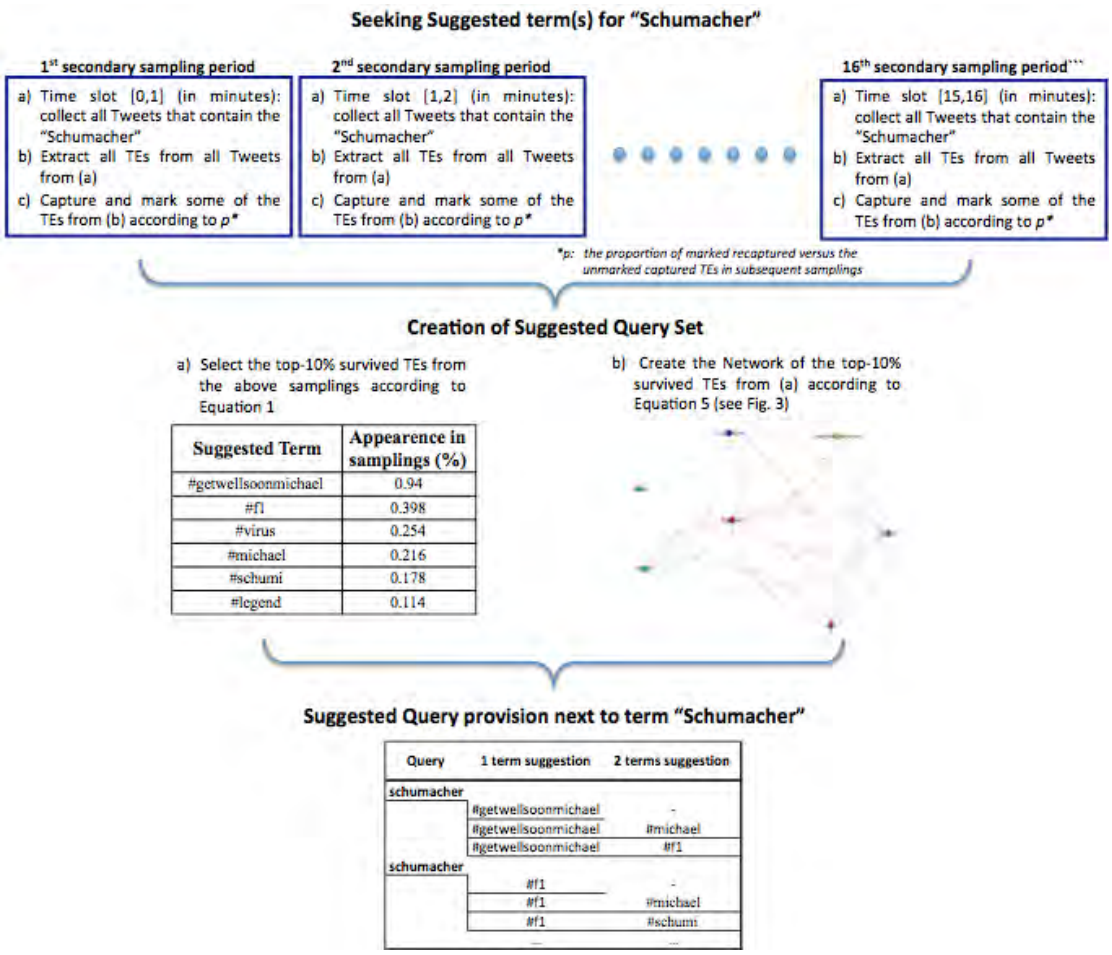

Figure 6.2: Flowchart of our proposed algorithm towards Query Suggestion provision with an example

We want to stress here that the capture-recapture paradigm helps us to not only suggest popular TEs, but also to suggest TEs that that remain popular in time. This practically means that we are not so interested in TEs that appear suddenly and then die out, but we in those that appear and re-appear in many subsequent sampling occasions. Thus, the concept of virality in this work is strongly related to survivability as mentioned in real-life capture-recapture experiments.

6.4.2. Weighting factor - ranking suggested terms

The main scope of the weighting factor is to calculate the weights of the most trending entities provided by the survivability factor (above sub-section) and then to provide their ranking position towards a query suggestion provision. After having the

top-k% most frequent (survived) entities according to the survivability factor of Equation 6.1, we further calculate their relation of co-appearance within the secondary sampling periods. This is performed by calculating their Twitter Semantic Weight (TSW) score according to Equation 6.2, where $ER(e_x, e_y)$ defines the frequency of co-appearance for entities x and y. This provides us with a ranking order (higher to lower TSW values) of the coupled Twitter Entities x and y.

$$TSW(e_x, e_y) = \frac{\|\varphi_i(\mathrm{e_x})\| + \left\|ER(e_x, e_y)\right\|}{2} \qquad (6.2)$$

Now, as far as our example is concerned, and as depicted in the flowchart of Figure 6.2, the network that consists of the top-10% survived TEs according to Equation 6.2 is illustrated in Figure 6.3. The most frequently appearing, i.e. viral entity, is the hashtag "#getwellsoonmichael" (TSW=0.448) and as a result it is proposed as the first suggestion term next to query "Schumacher", while the second one term suggestion is "#f1". Regarding the best two-term suggestions with respect to the same seed, these are "#getwellsoonmichael_#michael", as well as "#getwellsoonmichael_#f1", having TSW values equal to 0.181 and 0.177, respectively, followed by "#f1_#michael", "#f1_#schumi" with TSW equal to 0.084 and 0.082, respectively. The whole network that shows the relations and the respective weights between the survived entities in the query suggestion set of the Schumacher case, as well as the query suggestions, are shown in Figure 6.3.

We would like to note here that tokenization, topic/word segmentation, as well as further lexical analysis procedures that deal with breaking a stream of text into words/phrases are not considered in this study and are left for future work; it is obvious that they would be very useful, mainly for the hashtag entities.

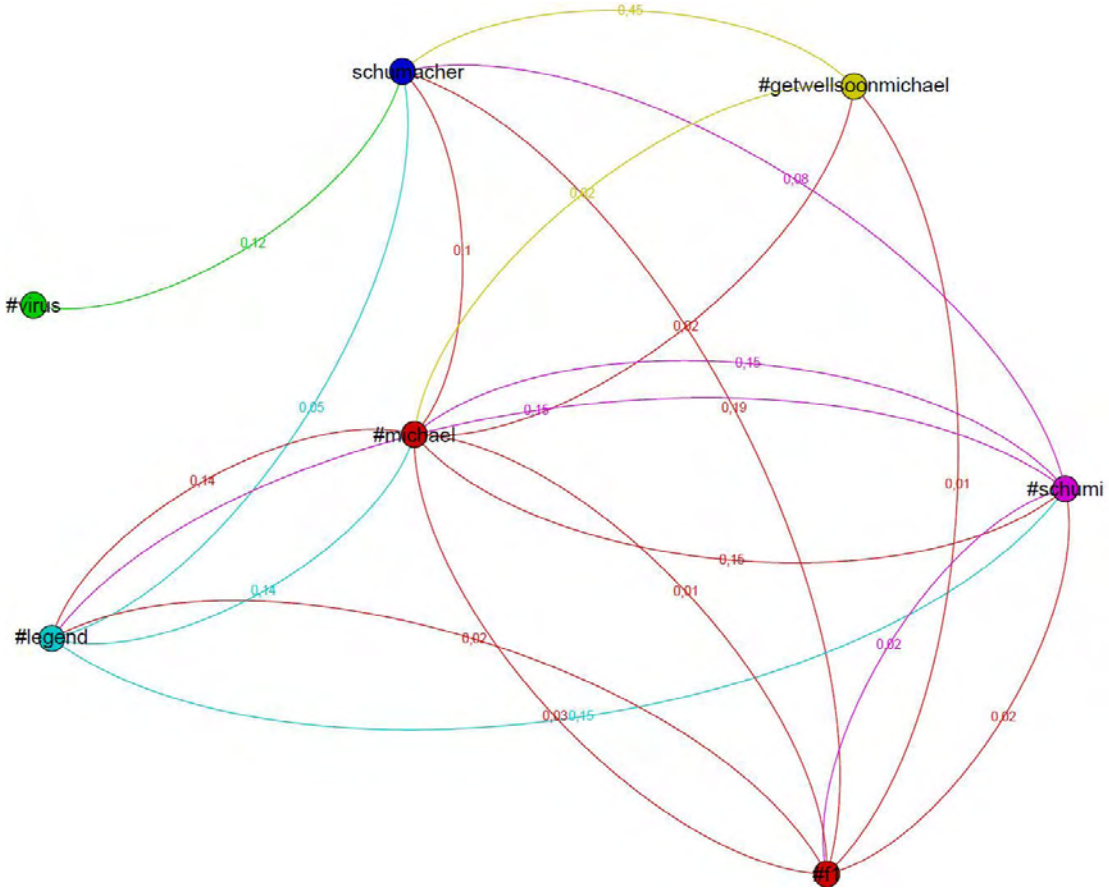

Figure 6.3: Network created from the top-10% survived TEs for the example provided in Section 6.4

6.5. Case studies: Gaining more insights in the suggested results

In this section we evaluate the results of the proposed query suggestion mechanism described in Section 6.4, by comparing the results derived from two case studies with respect to the query suggestions of well-known search engines, as well as to a heavily visited mainstream web media service As case studies in our experiments, we consider the political situations in Egypt and Syria, which have a constant interest for many years worldwide. Taking into account the knowledge we earned from our work (Anagnostopoulos et al., 2013), we initiated the query suggestion results procedure by using "Egypt", and "Syria" as seeds that correspond to our case studies. We divide the obtained results in two separate case studies. The experiments lasted for 8 days (January 8, 2014 to January 15, 2014). For each day, we conducted 8 primary sampling periods, while each primary period consisted of 16 secondary sampling periods, according to the modified capture-recapture model we followed (see Figure 6.1 and the rationale of the flowchart of Figure 6.2). In order to provide a thorough representation of the results, we analyze in detail the results of only one day (January 13th) for all the tested cases, and we then summarize them.

6.5.1. Query suggestion provision over the case studies

As mentioned before, we aim at providing query suggestions to the users' submitted term(s) under only the knowledge disseminated publicly in Twitter, and without having any other access or use of query logs.

Table 6.1 presents the entities with the top-k% survivability rate between 16 subsequent secondary samplings that appeared in January 13, 2014. The entities that reflect a significant trend behavior with respect to the seed term during that day (k=10) are thus highlighted. Unfortunately, no URL appeared as a survived entity among the top-10% population, while only two mention entities appeared - one for the case of "Egypt" and one for "Syria". The reader can see a few URLs and mention entities in Appendix D: Viral Twitter Entities for the cases of Egypt, and Syria, where *k* is equal to *30%*, and *20%*, respectively. Entities are depicted in alphabetical order and are colored differently according to their type (blue-highlighted: mentions, green-highlighted: hashtags, red-highlighted: URLs). For example, the entity @egypt_now111 presented the higher survivability during that day, since it was captured in nearly 88% of all the sampling occasions. Similarly, hashtag #freethe7 presented the higher survival rates for the Syria case, having a value equal to 83%. All the entities that are highlighted in Table 6.1a as well as in Table 6.1b form the Query Suggestion Sets for the two cases and they were derived according to the survivability factor φ_i as described in sub-section 6.4.1.

Table 6.1: Top 10%-survived TEs among 16 subsequent samplings of January 13, 2014 - (a) Egypt, (b) Syria

Query Suggestion Set (Seed: Egypt)	**Metrics**	
	TE type	***φ_i (top-10%)***
egypt_now111	@	0.881
anticoup	#	0.855
kuwait	#	0.615
saudi	#	0.602
morsi	#	0.539
uae	#	0.340
sta	#	0.168

(a)

Query Suggestion Set (Seed: Syria)	Metrics	
	TE type	φ_i *(top-10%)*
freethe7	#	0.830
iran	#	0.659
iraq	#	0.621
Free_Media_Hub	@	0.335
egypt	#	0.181
assad	#	0.160
un	#	0.143
isis	#	0.072

(b)

In conjunction to Table 6.1, Table 6.2 depicts some metrics among the top-3 survived entities and all the others that belong to the query suggestion set. These values are taken for all the three cases within the 16 subsequent sampling periods of January 13, 2014. In the third, fourth and fifth column of Table 6.2, we can see the frequency of co-appearances between entities x and y, as well between the seed term and x and y, respectively. For example, in the case of Egypt (Table 6.2a) hashtags #kuwait and #morsi appeared together in 324 captured tweets, while they co-appeared with the seed term (egypt) in 596 and 523 tweets, respectively.

Table 6.2: TSW parameters between the top-3 and the remaining survived entities (TEs) of the query suggestion set for January 13, 2014 - (a) Egypt, (b) Syria

seed:egypt				
entity(x)	entity(y)	freq.(x,y)	freq.(seed,x)	freq.(seed,y)
@egypt_now111	#saudi	0	854	584
@egypt_now111	#kuwait	0	854	596
@egypt_now111	#morsi	0	854	523
@egypt_now111	#uae	0	854	330
@egypt_now111	#anticoup	0	854	829
@egypt_now111	#sta	0	854	163
#anticoup	#saudi	0	829	584
#anticoup	#kuwait	0	829	596
#anticoup	@egypt_now111	0	829	854
#anticoup	#morsi	32	829	523
#anticoup	#uae	0	829	330
#anticoup	#sta	0	829	163
#kuwait	#saudi	560	596	584
#kuwait	@egypt_now111	0	596	854
#kuwait	#morsi	324	596	523
#kuwait	#uae	227	596	330
#kuwait	#anticoup	0	596	829
#kuwait	#sta	163	596	163

(a)

seed:Syria				
entity(x)	entity(y)	freq.(x,y)	freq.(seed,x)	freq.(seed,y)
#freethe7	#isis	0	545	47
#freethe7	#egypt	15	545	119
#freethe7	@Free_Media_Hub	0	545	220
#freethe7	#iraq	124	545	408
#freethe7	#iran	220	545	433
#freethe7	#un	183	545	94
#freethe7	#assad	0	545	105
#iran	#isis	1	433	47
#iran	#egypt	9	433	119
#iran	@Free_Media_Hub	0	433	220
#iran	#iraq	142	433	408
#iran	#freethe7	220	433	545
#iran	#un	198	433	94
#iran	#assad	0	433	105
#iraq	#isis	6	408	47
#iraq	#egypt	8	408	119
#iraq	@Free_Media_Hub	0	408	220
#iraq	#iran	142	408	433
#iraq	#freethe7	124	408	545
#iraq	#un	141	408	94
#iraq	#assad	6	408	105

(b)

Taking into account ER(e_x, e_y), the algorithm dynamically calculates the weights according to Equation 6.2 and generates the network of related entities for further query suggestions. Figures Figure 6.4a, and Figure 6.4b depict these networks with respect to the cases we investigate. The nodes correspond to the entities of the query suggestion set (survived entities from the capture-recapture experiments), while the curving edges indicate clockwise the direction from a source to the target node and the respective weight. The networks (layout type: Fruchterman Reingold) are created with the open graph visualization Gephi tool.

In Appendix D: Viral Twitter Entities, we provide as many results as they can practically be depicted for these two cases. As we have mentioned earlier, the reader can see the query suggestion sets with all the entities which have the top-k% survival rates within the daily capture-recapture experiments ("Egypt": k=30%, "Syria": k=20%).

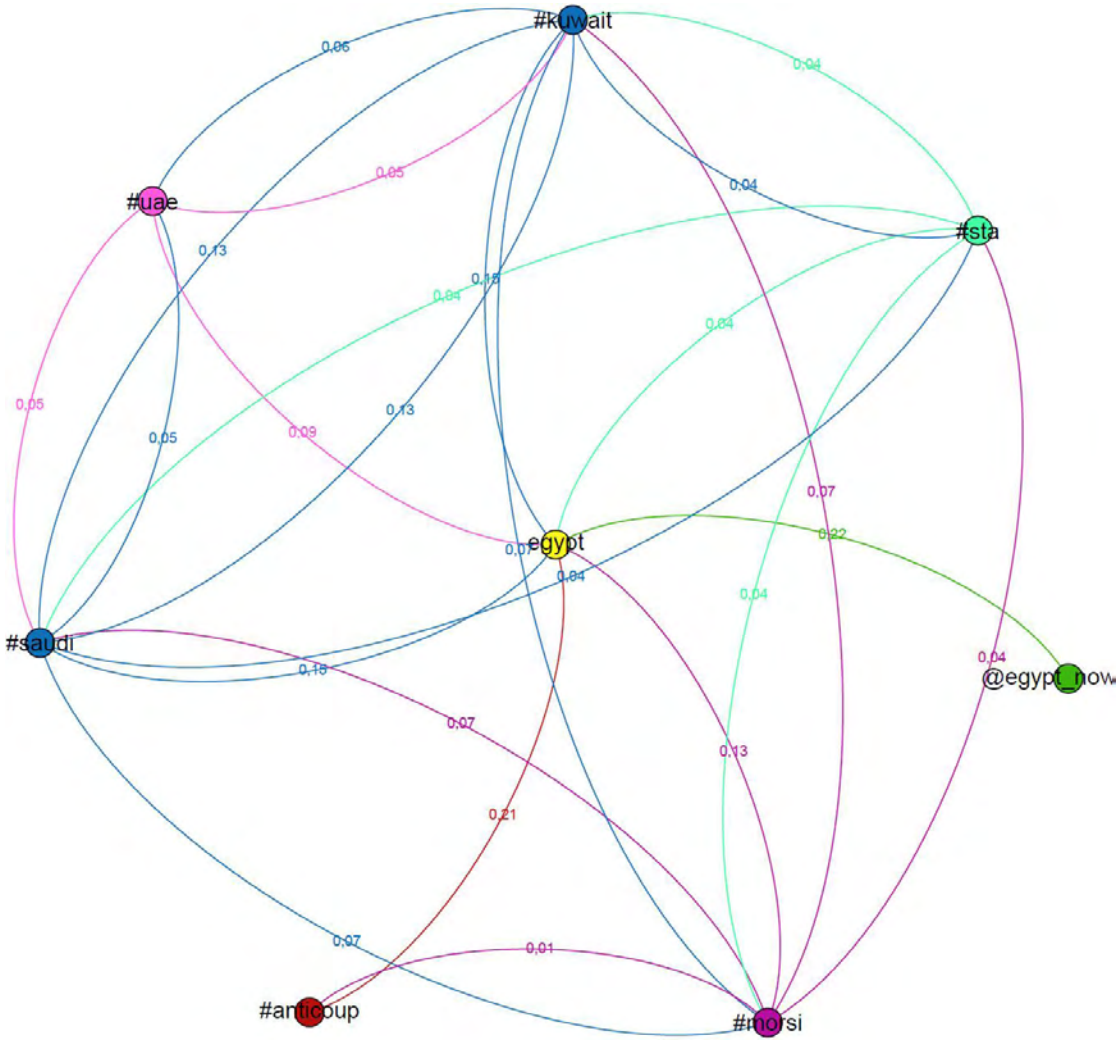

Figure 6.4a: Network of survived Twitter Entities in Query Suggestion Sets - Egypt

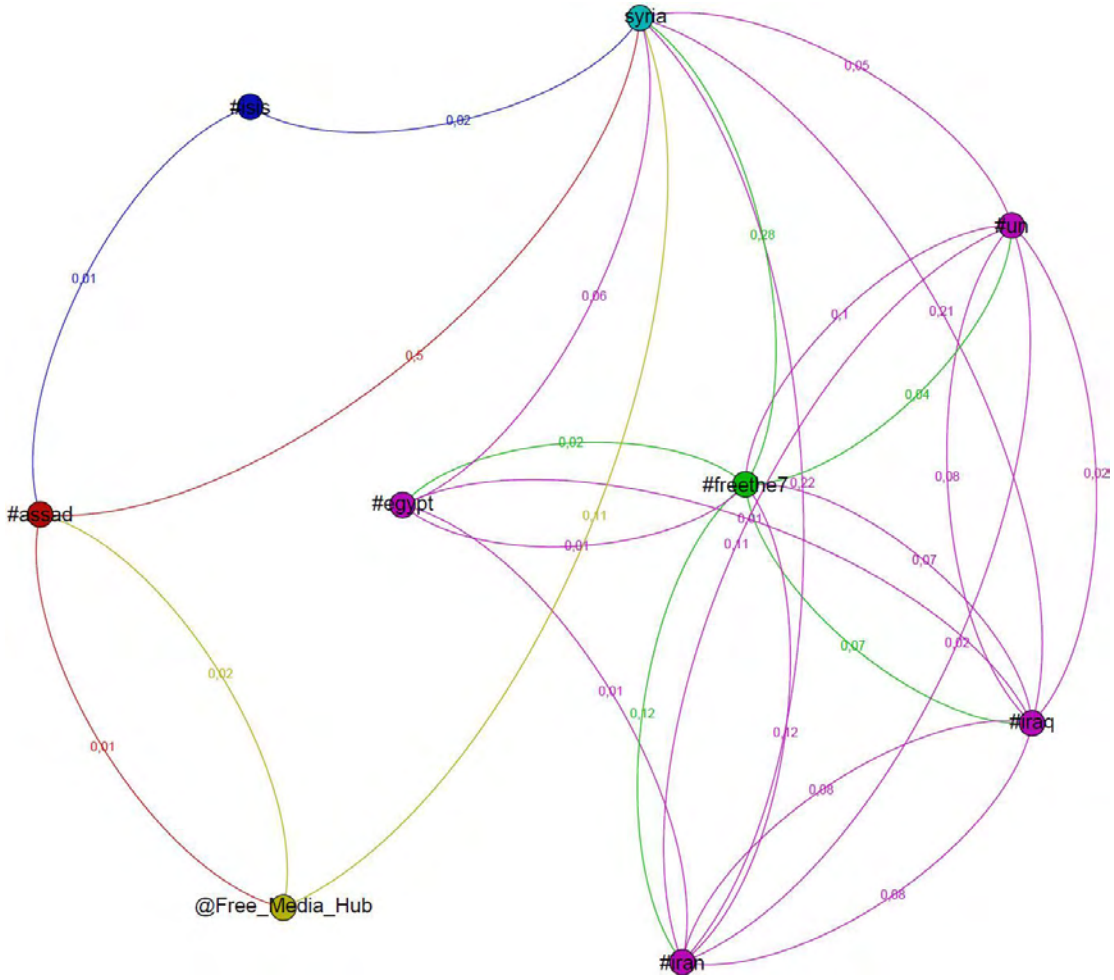

Figure 6.5b: Network of survived Twitter Entities in Query Suggestion Sets - Syria

6.5.2. Evaluation of results against major search services

In this subsection we evaluate the query suggestions provided by our approach by comparing the query recommendations with respect to the recommendations of Google News[28], Yahoo! News[29], and Bing[30], as well as to the Reuters portal[31] for the two cases (seeds: "Egypt", "Syria"). We should note here that our intention is not only to evaluate the accuracy of the provided suggestions, but also to investigate how quickly the suggestions reflect their trends on a daily-basis. Our query suggestion set consists of user-generated content (#hashtags), main influencers and web content (Twitter accounts, mentions, URLs), which is disseminated and semantically annotated in Twittersphere. The innovation of our method relies in the network analysis of a large ecosystem that involves users, semantics, and web content and not from query logs. In other words, suggestions are not driven from past user searches, but are performed nearly on-the-fly and directly from the Twitter API. According to our knowledge, Google's predicting algorithm used for query suggestion displays search queries based on other users' search activities and the contents of Web pages

[28] https://news.google.com/
[29] http://news.yahoo.com/
[30] http://www.bing.com/news/
[31] http://www.reuters.com/

indexed by Google[32]. In addition, Google users might also see search queries derived from their previous related searches. We suppose that the other information services we use for comparison (Yahoo! News, Bing and Reuters) practically work under the same concepts. Still, we should note at this point, that if the search service uses a *search results* based approach, the query suggestion depends mainly on a specific number of the top-N results for that query. Yet, if the service uses *logs*, the query suggestion may be provided using other relevant user query terms, or even other user personalized behavior-based search pattern.

Prior to starting our evaluation and discussion regarding our results, we introduce Table 6.3, which presents the query suggestions provided by Google, Yahoo!, Bing and Reuters within our testing period between the 8th and the 15th of January 2014 for the two case studies. This table is divided into four parts that reflect the results of the above-mentioned news services with respect to the tested seed terms. We can notice that Google and Yahoo! provided 10 suggestions per seed, while Bing and Reuters 9 and 4, respectively. In the first column of each part we have the ranking position of the suggestion. For example, "constitution" was ranked as the third suggested term (seed: "Egypt") from Google, and "protests" as the fifth suggested term (seed: "Syria") from Yahoo!. In addition, the blue highlighted terms followed by an asterisk denote terms the rank of which interchanged during the testing period. For example, in Bing results and for the seed "Egypt", the terms "map" and "fact" firstly appeared as the 6th and the 8th result, respectively, but later on their ranking positions were switched. Finally, the green highlighted terms followed by "+", denote new terms that appeared on the position of an already suggested one. This means that in such cases, the suggested terms are updated and the list is refreshed. For the period of our tests, such refresh activity appeared mainly from Reuters and less for Bing. For example, "iran revolution" was the fourth suggested term from Reuters with respect to the seed "Egypt" for January 8 and 9. Then, the next day (January 10th), it was replaced in the same ranking position by the term "qatar", then by the term "egyptian" (January 11) and then by the term "constitution" for the remaining days (January 14-15). Similarly for the seed "Syria", new suggestions over "the snipers of" appeared on the fourth ranking position during January 12, 13 and 14 with the two-term suggestions "latest news", "Syrian war" and "gas attacks", respectively. Regarding the other three search engines, only Bing provided updated content, where the term "newspaper" was replaced by the term "snow" on January 14th.

[32] http://support.google.com/websearch/bin/answer.py?hl=en&answer=106230

Table 6.3: Query suggestions provided by Google, Yahoo!, Bing and Reuters with respect to our case studies (January 8 - January 15 2014)

Google News			Yahoo! News		
rank	egypt 8-15 Jan 2014	syria 8-15 Jan 2014	rank	egypt 8-15 Jan 2014	syria 8-15 Jan 2014
1	snow	news	1	protests*	news
2	news	chemical weapons	2	news*	war 2013
3	constitution	rebels	3	shark attack	chemical attack
4	turkey	kurds	4	pyramids warning	chemical weapons
5	protest	nuns	5	elections	protests
6	economy	aleppo	6	locusts	israel
7	ghana	chemical	7	crisis 2011	uprising
8	russia	war	8	__'s president	russia
9	referendum	children	9	antiquities	turkey
10	clashes	fighting	10	israel	rebels

Bing News			Reuters		
rank	egypt 8-15 Jan 2014	syria 8-15 Jan 2014	rank	egypt 8-15 Jan 2014	syria 8-15 Jan 2014
1	news	news	1	news	the syrian front*
2	newspaper	latest news	2	economy*	news*
3	*snow (14 Jan)*		3	reuters*	chemical weapons
4	air	chemical weapons*	4	iran revolution	the snipers of
5	sherrod	map		*qatar + (10 Jan)*	*latest news + (12 Jan)*
6	map*	now*		*egyptian + (11 Jan)*	*syrian war + (13 Jan)*
7	chaos	war		*constitution + (14 Jan)*	*gas attacks + (14 Jan)*
8	facts*	tv*	*terms that their rank interchanges		
9	pyramids	tube	+new entry (date)		

The evaluation results with respect to these cases and our method are presented in Table 6.4, where we can see on the left the initial query terms (seeds), followed by the suggested entities in two levels, corresponding to entity(x) and entity(y). These levels resemble to the automatic recommendation provided by several search engines based on the already submitted user term(s). Finally, column "TSW weights" corresponds to the weighting values for the query suggestions "{seed}- _entity(x)" and "{seed}_entity(x)_entity(y)". For simplicity reasons, Table 6.4 holds only the TSW values of the first two second level entities *y* that are suggested by entity *x*.

Table 6.4: Query suggestions according to Twitter Semantic Weighting (January 13, 2014) - (a) Egypt, (b) Syria

seed	1st level suggestion entity(x)	2nd level suggestion entity(y)	*TSW weights*	
egypt				
	@egypt_now111		*0.220*	
	#anticoup		*0.214*	
		#morsi		*0.136*
	#kuwait		*0.154*	
		#saudi		*0.117*
		#morsi		*0.108*
	#saudi		*0.151*	
		#kuwait		*0.115*
		#morsi		*0.106*
	#morsi		*0.135*	
		#saudi		*0.097*
		#kuwait		*0.097*
	#uae		*0.085*	
		#kuwait		*0.063*
		#saudi		*0.062*
	#sta		*0.042*	
		#kuwait		*0.033*
		#saudi		*0.033*
		#morsi		0.033

(a)

seed	1st level suggestion entity(x)	2nd level suggestion entity(y)	*TSW weights*	
syria				
	#freethe7		*0.277*	
		#iran		*0.147*
		#un		*0.139*
	#iran		*0.220*	
		#freethe7		*0.126*
		#un		*0.121*
	#iraq		*0.207*	
		#iran		*0.105*
		#freethe7		*0.105*
	@Free_Media_Hub		*0.112*	
		#assad		*0.046*
	#egypt		*0.060*	
		#freethe7		*0.029*
		#iraq		*0.026*
		#iran		*0.026*
	#assad		*0.053*	
		@Free_Media_Hub		*0.026*
		#isis		*0.023*
	#un		*0.048*	
		#freethe7		*0.032*
		#iraq		*0.025*
		#iran		*0.024*
	#isis		*0.024*	

(b)

6.5.2.1. Egypt case

Moreover, for the case of political situation in Egypt, the entity that presented the higher TSW during the 13th of January 2014 (equal to 0.220) was the Twitter account @egypt_now111. This account disseminates breaking news regarding the unstable political situation in Egypt, having nearly 170,000 followers and 19,000 tweets per day. It is worth noticing that Al Jazeera's Twitter account has less than half of the followers and tweets (nearly 65,000 and 8,000, respectively). Even though there were many other Twitter Entities captured and marked along with @egypt_now111, there is no second level suggestion since most of them are written in Arabic. That is why in Figure 6.4a there is no other edge from the @egypt_now111 node to the other nodes of the query suggestion set. The entity with the second higher TSW (0.214) is the hashtag "#anticoup", which obviously comes from combining the Greek term "anti-" (expressing opposing to or against to something/someone) and "coup". This proposal was highly relevant and quite trendy with respect to the political status within the testing period, since there were demonstrations in many Egyptian cities and villages condemning coup crimes. This trend was not captured by the other services as we can see from Table 6.3. However, the suggestion "protest(s)" that reflects similar actions over the coup in Egypt, was suggested by Yahoo! and Google in the first and fifth place accordingly. As a second level suggestion related to "#anticoup" is "#morsi", having a TSW equal to 0.136. This means that in case we want to find two-term suggestions (entities in this paper) for our seed, then the sequence "#anticoup #morsi" is the most frequent.

Then, as the third, fourth and fifth recommendation with respect to the seed "Egypt", we have the hashtags "#kuwait", "#saudi" and "#morsi" which are strongly related to each other, since all are suggested by the other two in a second level. For instance, "#saudi" and "#morsi" suggest "#kuwait", "#kuwait" and "#morsi" suggest "#saudi", as well as "#saudi" and "#kuwait" suggest "#morsi" according to a descending weighting order. This can be also seen in Figure 6.4a where we can notice that the corresponding entity nodes are networked one-by-one. This strong network relation between these three entities is justified, since during the period of our experiments the trial of the ex-president of Egypt (Morsi) was supposed to take place on January 8 (eventually it was postponed). In addition, a day prior to this trial there was a political declaration in favor of the new Egyptian government and against Morsi's believers by Saudi Arabia, the United Arab Emirates, and Kuwait. This was also noticed with the "#uae" as the sixth first level suggestion, where even though the TSW values were lower, yet "#uae" was suggested by "#kuwait" and "#saudi" at a second level. The whole network that shows the relations and the respective TSWs in the query suggestion set of seed "Egypt" case, as well as the proposed query suggestions, are shown in Figure 6.4a and Table 6.4a, respectively.

6.5.2.2. Syria case

Results with respect to "Syria" as a seed term revealed strong relations between the entities "#freethe7", "#iran", "#iraq", and "#egypt". Similarly with the above explanation, the top-3 suggestions included "#freethe7", "#iran", "#iraq" with TSW values 0.277, 0.22 and 0.207, respectively. Regarding the top two-term suggestions with respect to the seed, these were "syria_#freethe7_#iran", "syria_#freethe7_#un" with TSW values equal to 0.147 and 0.139, respectively, followed by

"syria_#iran_#freethe7" and "syria_#iran_#un" with TSW values of 0.126, 0.121, respectively. The trending behavior of these entities is justified not only due to the political positions and status between those countries over the last years, but also because a couple of days before capturing these trends, UN experts urged Iraq to establish once again the fate and whereabouts of the seven residents of Camp Ashraf, who were allegedly abducted on September of 2013 after an attack in which more than 50 persons were killed. The network that shows all the relations and their respective weights between the proposed entities in this case, as well as the query suggestions, is shown in Figure 6.4b and Table 6.4b, respectively.

6.5.2.3. Comparison with other services

After observing the query suggestions provided by the other search services (see Table 6.3), it surely worth discussing the performance of Google, Yahoo!, Bing and Reuters in terms of query suggestion freshness (how often these services update their suggestions after a query). Google News surprised us negatively since its query suggestions were identical and static for all the evaluation period (from January 8, 2014 up to January 15, 2014). Yahoo! News had also static suggestions with respect to the tested terms. There were only some re-rankings for the first two proposed suggestions in the case of Egypt, but no new entries. On the other hand, Bing presented more freshness activity mostly related to the re-ranking of its suggestions. In addition, nearly the end of our evaluation period (January 14), Bing replaced the suggested term "newspaper" with the term "snow" in the second position (seed "Egypt"). However, this replacement was quite outdated since it was related to a snowfall in the Northern Egypt territories, which happened nearly a month before (in mid December of 2013).

Among all, Reuters provided the most dynamic results in terms of freshness and position re-ranking. Despite the fact that Reuters returns fewer suggestions, these are updated quite often. From Table 6.3 we notice that within an eight-day period and for the cases of Egypt and Syria, the last suggestions were replaced by newer ones at least three times (10^{th}, 11^{th}, 14^{th} and 12^{th}, 13^{th}, 14^{th} of January, respectively). According to our research in query suggestion for web news services, this is the second time where Reuters receives the best comments with respect to trendy proposed suggestions near a user's query (see the evaluation described in (Anagnostopoulos et al., 2012)). Finally, in an attempt to directly compare if our approach "captures" the new replacements made by Reuters, we checked whether these suggestions appeared in our records. So, as it can be seen in Appendix D: Viral Twitter Entities, the entities "#qatar", "#egyconstitution" (obviously concatenation of "Egypt" and "constitution"), as well as the entity "#gasattacks" appeared among the top-30% and top-20% survived entities with respect to the cases of Egypt and Syria.

6.6. Evaluation - Discussion

The overall evaluation of our proposed methodology follows two lines. At first, we describe a generic evaluation, which involves subjective user ratings for results obtained from our approach and from Google Hot Trends. Then, we provide comparative evaluations with respect to two similar baseline methods from the

literature, namely a cluster labeling and a microblog retrieval task, over traditional information retrieval metrics.

6.6.1. Evaluation against user ratings

In order to evaluate whether query suggestions returned from our method satisfy the user's information needs we engaged 17 postgraduate students from an MBA course class at the National Technical University of Athens. Their task was to subjectively rate the suggested queries against the Google Hot Trends service[33]. Google Hot Trends displays several top fastest rising searches (and search-terms) by day in the U.S.A. Each student was asked to select three different events from Google Hot Trends for a specific testing period. Furthermore, each individual had to explicitly rate the suggested entities as these were derived by our query suggestion method, against their selected events as appear in Google Hot Trends. The rating performed upon a five-point Likert scale (see Table 6.5).

For simplicity reasons and without any loss of generality, the students were asked to rate only hashtags as extended entities (terms) and considered only the survivability factor (as described in sub-section 6.4.1). After processing the one-week results we ended up with 87 unique related terms (as these were provided by Google Hot Trends) in 31 distinct events (20 out of the 51 events were identical). The average number of suggested terms per tested event was 2.81, which practically means that nearly 3 terms on the average suggest the basic term that describes a specific event. The inter-annotator agreement was the Fleiss' kappa statistical measure for assessing the reliability of agreement between a fixed number of raters. In our evaluation we had 17 individuals (raters) for assigning 87 unique related terms (subjects) to a total of 31 distinct events. The value of kappa was measured at nearly 0.37, which is an almost fair agreement according to the literature (Geertzen, 2012). Figure 6.6, we can see some points that indicate the average evaluator rating for suggested entities, as derived from our proposal. In addition, Table 6.6 summarizes all the mean rate values with respect to the survival rate, as well as other parameters taken were into consideration in this evaluation.

Table 6.5: The five-point Likert scale used for user ratings

1	2	3	4	5
Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
(totally irrelevant suggestion)	(not so good suggestion)	(nearly same suggestion)	(potentially better)	(surely better)

[33] http://www.google.com/trends/hottrends

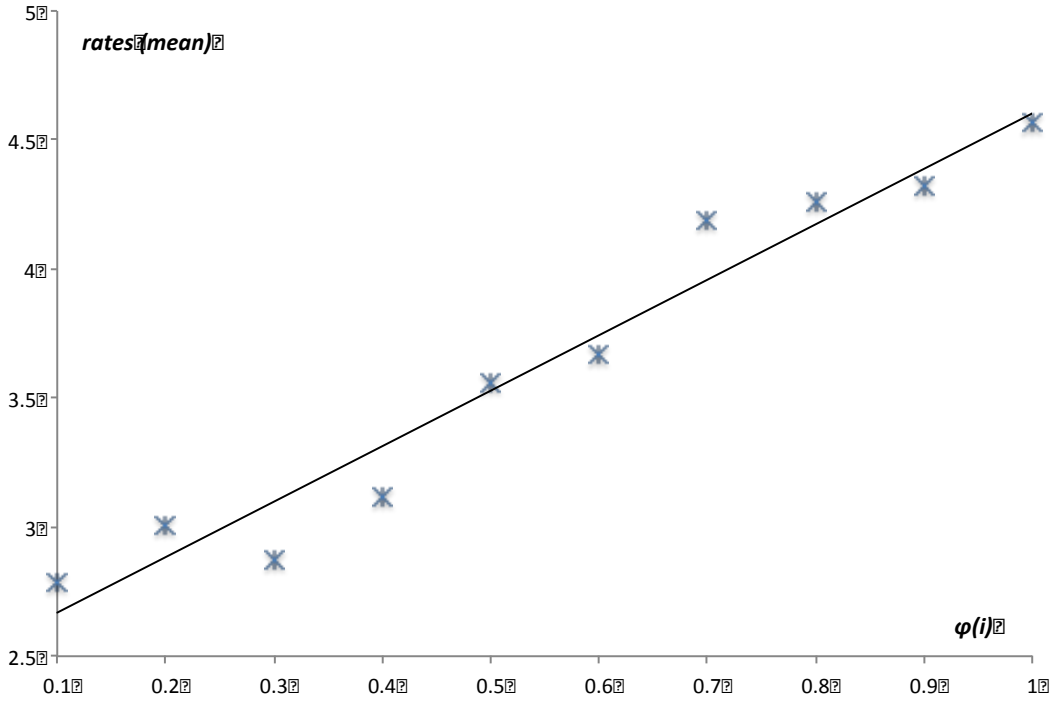

Figure 6.6: Rates (in mean values) versus proposed Twitter Entities from Query Suggestion sets

Table 6.6: Evaluation metrics against user ratings

φ(i)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
rates (mean values)	2.72	3.05	2.79	3.12	3.55	3.61	4.21	4.34	4.41	4.48
# of individuals (raters): *17*										
unique related terms (subjects): *87*										
distinct events: *31*										
inter-annotator agreement: *Fleiss' kappa statistical measure, k = 0.37*										

It is worth noticing that the larger the survival rate (φ), the higher the mean subjective rate appears in the five-point Likert scale. This proves that in this way the query suggestion set formed, consists of more trendy entities related to each other. More specifically, we noticed that entities belonging in query suggestion sets that had survivability rate above 0.7 were subjectively evaluated as more relevant, as they presented nearly one-point higher level in the Likert scale. This practically proves that through subsequent samplings in Twitter, the most viral entities are trendier in comparison to a related query term residing in Google's log. This was somehow expected, since we performed a short-term trend analysis rather than a long-term log analysis, yet it is an indicative assumption that our query suggestion method is in the right direction. We can also notice that the majority of the subjective rates in average values (more than 60%) were close or slightly higher in the third level (point 3) in the

Likert scale, thus indicating a "nearly same suggestion" in comparison to the compared Google search service.

6.6.2. Evaluation against a cluster labeling and a microblog retrieval task

Towards a more comprehensive evaluation, we compared our approach with two similar baseline approaches, namely a cluster labeling and a microblog retrieval task.

6.6.2.1. Comparative evaluation against a cluster labeling task

During the first comparison approach we consider as baseline the evaluations taken into consideration in the work described in (Hu et al., 2011). In this work, the authors propose methods to aggregate related microblogging messages into clusters and automatically assign them semantically meaningful labels. They use hot queries of diverse topics selected from Google Hot Trends, where each query is considered to be a trending topic, while they consider the top-five query suggestions from Google as subtopics of this topic. In order to be able to compare our approach on a common basis, we considered the suggested TEs to form a cluster similar to the ones in (Hu et al., 2011). This means that the generation of relevant labels around a topic/subtopic is considered similar to selecting appropriate terms in the Query Suggestion Set of our method. So, for collecting the pool of our data, we systematically crawled the hot queries published by Google Hot Trends between May 9 and June 9 of 2014, having in mind 20 hot queries from miscellaneous fields of interest, for which Google provided some relevant/similar queries. Each selected hot query was considered as a different topic. Then, for each topic we further crawled the top-three query suggestions, thus forming 60 separate subtopics. The returned suggestions formed a cluster, while the subtopics were considered associated with the cluster label. For each subtopic (query suggestion), we harvested exactly the last 200 tweets from Twitter. As a result, we collected nearly 12,000 tweets with all of their TEs.

The next step included the comparison of our clusters with the clusters formed by the three methods mentioned in (Hu et al., 2011), namely the WordNet_Method (WNT), the Wiki_Method (WK) and the SemKnow_Method (SMK). We used the harmonic mean of precision and recall (F1-score) and the Accuracy metrics to evaluate the performance of the compared methods. Table 6.7 depicts the values of these metrics for all three methods, as well as ours. Since the clusters in our method dynamically change every three hours (time interval between subsequent primary periods), the respective values in the last row of Table 6.7 correspond to their average values across the entire testing period. We notice that all approaches improve the classical BOW model, both in terms of F1-score and Accuracy. Still, the best improvement appeared for our method and it is 13.6% better with respect to the BOW model and 6.5% better compared (with respect to absolute F1-score values) to the second best method (SMK). In terms of Accuracy, our method presented a 16.8% and a 9.8% (with respect to absolute Accuracy values) improvement having BOW and SMK as baselines, respectively. Similarly to the work described in (Hu et al., 2011) we notice that SMK, WK and WNT increase (in that particular order) the values of the used metrics.

Table 6.7: Results from methods that integrate the BOW and our method (cluster labeling task)

Method / Metric	F1-score	Deviation over BOW (±%)	Accuracy	Deviation over BOW (±%)
BOW	0.468	N/A	0.510	N/A
WordNet_Method	0.476	+1.735	0.520	+1.919
Wiki_Method	0.499	+6.687	0.541	+6.023
SemKnow_Method	0.500	+6.929	0.543	+6.477
Our Method	**0.532**	**+13.624**	**0.596**	**+16.825**

In order to enhance the aforementioned evaluation, we considered the suggested top-5 TEs generated by our weighting factor (Equation 6.2) to be the best labels per subtopic for our method. This allowed us to directly compare our method with WNT, WK and SMK, treating the cluster-labeling task as a problem that ranks all the concepts from Wikipedia and the best matched label for a cluster of microblogging messages. As ground truth for cluster labeling, we considered the subtopics used for crawling microblogging messages. Table 6.8 summarizes the results based on the normalized Discounted Cumulative Gain up to the fifth position of the ranked results (nDCG5). Similarly to the previous described clustering evaluation, the values in the last row of Table 6.8 (Our Method) correspond to averaged values across the testing period. As we can see, our method presents the best normalized Discounted Cumulative Gain value in comparison to the other baselines. We strongly believe that this is due to the fact that our method provides quite up-to-date and "fresh" query suggestions in terms of TEs (Hashtags, User mentions, URLs) that derive directly from the users' intelligence and capability to describe content.

Table 6.8: Results from methods based on nDCG5 and our method (ranking problem)

Method / Metric	$nDCG_5$	Deviation over Kphrase (±%)
Kphrase	0.438	N/A
WordNet_Method	0.448	+2.281
Wiki_Method	0.520	+18.761
Our Method	**0.576**	**+31.484**

6.6.2.2. Comparative evaluation against a microblog retrieval task

A second evaluation procedure was performed against two other baselines described in (Massoudi et al., 2011), where the authors present a model for retrieving microblog posts enhanced with textual and microblog quality indicators, as well as with a dynamic query expansion model. In particular, we wanted to test the ability of

our query suggestion mechanism in terms of viral terms recommendation given a trending topic. So, for the same period utilized in our previous evaluation (May 9, 2014 to June 9, 2014), we selected some trending topics as proposed by Twitter, thus forming 20 different queries. Working similarly to (Massoudi et al., 2011), we harvested all the tweets that were posted between the very last day the topic was announced as trending and three days before that day, ending up with nearly 28,400 tweets. We then followed a simple procedure that required retrieval experiments with respect to the top-5 results for all the trending topics that fall within the three-day time window. We should note here that if a topic presented a trending behavior for more than one day (and there were many such cases!), the experimentation ran only for the first three-day time window, just before the day the topic appeared in Twitter Trends for the first time. The results were judged as relevant or not. The inter-annotator agreement was once more the Fleiss' kappa statistical measure for assessing the reliability of agreement between the raters, who in this case are equal to 4. The value of the kappa statistical measure was measured at the level of 0.74, while the evaluation metrics used were the Mean Reciprocal Rank (MRR), the Mean Average Precision (MAP), and the Precision at the fifth position (P@5). The baseline was a boolean search method, strongly biased towards newer results. That means that newer tweets were ranked in higher position. This baseline is called "*Boolean search with recency features*" (BS+R). Joint to this method, a classical relevance model (RM2) was also employed (Lavrenko and Croft, 2001).Table 6.9 depicts the metric values we considered for evaluation purposes, as well as the deviation over the baseline. We observe that RM2 improves the Mean Average Precision as well as the Precision at the fifth position by nearly 8.8% and 5.3%, respectively. For the same method, the Mean Reciprocal Rank was measured in nearly 15% lower level values with respect to BS+R. Now as far as our method is concerned, it performed significantly better than the rest, considering all the above metrics; the Mean Reciprocal Rank was improved by 27.7% and nearly 50% (in terms of absolute MRR values) compared to the baseline and RM2, respectively. Similar improvements were achieved for the Mean Average Precision values at the levels of 31.8% and nearly 21% (in terms of absolute MAP values), respectively. However, the most impressive improvement was measured for the P@5 metric. Our approach returned more than 91% higher precision level in comparison to BS+R (reaching the value of 0.9), while it also outperformed the traditional RM2 method by leveraging the P@5 value up to nearly 82% (in terms of absolute P@5 values). This actually proves how significant it is to provide viral suggested terms (in our case TEs) through a query suggestion method and not only rely on recency criteria. In addition, this kind of evaluation revealed that tokens with numeric or non-alphabetic characters, which are usually eliminated by traditional information retrieval methods, are of great importance towards query suggestion in microblog post search.

Table 6.9: Evaluation metrics for BS+R, RM2 and our method

Method / Metric	MAP {Deviation over BS+R (±%)}	MRR {Deviation over BS+R (±%)}	P@5 {Deviation over BS+R (±%)}
BS+R	0.362 {N/A}	0.723 {N/A}	0.468 {N/A}
RM2	0.394 {+8.831}	0.615 {-14.938}	0.493 {+5.341}
Our Method	**0.477 {+31.768}**	**0.923 {+27.663}**	**0.897 {+91.667}**

Chapter 7. Conclusions - Future Work

In this thesis, we presented novel models, methods and experimental results that focus on three major directions of OSNs, namely i) *social influence*, ii) *social semantics*, and iii) *qualitative assessment*.

In Chapter 2 we conducted a review covering two major aspects of OSNs, namely the online social influence and the role of semantics, while discussing how we can combine both aspects towards the qualitative assessment and modeling of user-generated content. We presented in details the methodologies as described in the most up-to-date and impactful studies relevant to the aforementioned aspects. Specifically, we examined what kinds of methodologies are used to measure influence and the factors they are based on, along with the application domains. Moreover, we analyzed works based on Semantic Web technologies along with network theory and graph properties for identification of topics, similar users and communities, as well as for user personalization.

In Chapter 3 and Chapter 4 we presented "*InfluenceTracker*"[8], a publicly available website[8] where anyone can rate and compare the recent activity of any Twitter account. The core of this service is "*Influence Metric*", presented in Chapter 3, a novel metric aiming at calculating the importance and impact of Twitter accounts, which is derived from a social function incorporating the activity of an account, its social degree and its qualitative content. Moreover, we introduced a new qualitative factor based on the established h-index metric. Its aim is to reflect other users' actions and preferences (i.e. retweets and favorites) over the content and properties of viral posts, thus enhancing the "Influence Metric" of real influencers. The conducted experimental evaluation provides evidence that the proposed methods improve the state of the art and are scalable to large amounts of data. Moreover, we proposed a methodology regarding the maximization of diffusion of information in OSNs. Finally, we performed extensive experimental evaluations against real-data use cases. The results show that the number of followers an account has, is not sufficient to guarantee the maximum diffusion of information in Twitter, and practically to any similar OSN. This is because, the Twitter accounts should not only be active, but also have an impact on the network. The latter is calculated by the "Influence Metric". As a future work, we plan to combine the InfluenceTracker metrics of "Daily h-indexes", representing the quality of the disseminated content, with the "Reply Percentage", indicating the conversational tendency of accounts, in order for spam accounts to be discovered.

In Chapter 4, we defined the "*InfluenceTracker Ontology*"[13], an easily extensible to other OSNs schema, for transforming unstructured social data, such as accounts, metadata, social relationships, entities (mentions, replies, hashtags, photos, and URLs), as well as other social and qualitative metrics into Linked Data. In addition, we provided details on how this semantified information can be linked to the LOD cloud, thus increasing the value of the available data. The structured and semantified information is publicly available for querying through the provided SPARQL endpoint[10,11] of the *InfluenceTracker*[8] service, where a federated query demonstrates the benefits of the semantic technologies and the LOD cloud by providing answers to sophisticated queries (e.g. return the top-10 members of political parties according to their "Influence Metric" value). To the best of our knowledge, there is currently no active service for providing such kind of data linkage, i.e. social analytics with the

LOD cloud. Since the latest update of the LOD cloud[16], on 20/02/2017, the "InfluenceTracker" dataset is officially part of this interlinked and interdependent ecosystem of data. Finally, we proposed and analyzed two generic and adaptable methodologies for discovering the linked data resources best describing Twitter entities for further enhancing the quality of existing information. In the future, we plan to expand our service by incorporating data from other ontological schemes found in the LOD cloud (e.g. MusicBrainz[34], LinkedMDB[35], FlickrWrappr[36]) in order to cover more aspects and to provide answers to more sophisticated and complicated queries.

In Chapter 5 we described methodologies adopting Semantic Web technologies towards user classification, topic identification and data enrichment. Specifically, we proposed a framework for suggesting similar accounts to Twitter users by considering their common disseminated content and relations in Twittersphere. For defining the similarity metrics we exclusively employed Semantic Web technologies and models (e.g. see SPARQL Queries 1 and 2 in Appendix C: SPARQL Queries) based on the proposed *InfluenceTracker Ontology* (Section 4.3.2). The existence of an ontological scheme and the use of semantic technologies reduced the complexity of storing and retrieving specific segments of data and decreased the number of the necessary calculations required for the computation of the proposed algorithm's coefficients and metrics. The results of the conducted evaluation, involving subjective user ratings, showed that the majority of those rates (in average values) were very satisfying. Moreover, based on the similarity algorithm, we developed a methodology towards the automatic labeling of Twitter accounts with respect to thematic categories derived from DBpedia properties, in order to classify them into communities. That methodology demonstrates the joint benefits of the semantic technologies and the LOD cloud towards the enrichment of Twitter data with information from other sources. To the best of our knowledge, this is the first work and public service which combines semantified social analytics with the LOD cloud. In the future, we plan to extend those methodologies for highlighting communities (of different sizes) of Twitter accounts of similar content, influence and activities. It is also worth investigating the dynamics of such communities across different thematic domain and real-time events. Finally, by adjusting our proposed algorithms the identified communities can be accompanied by tag clouds around them enriched with their most distinctive social information (i.e. mentions, replies, hashtags, photos, URLs).

In Chapter 6, we introduced a query suggestion method based on a social network, derived by related trendy entities that become viral in Twitter worldwide. The innovation in this work stands in the fact that we used the users' intelligence and capability to describe information (e.g. through the hashtags), as well as the power that social media have to validate enhance, or modify it in real-time. In comparison to other query suggestion methods, the added value of our proposal is two-fold, since we achieve better freshness and trendiness rates for our query suggestion set. We witnessed many cases in which suggestions proposed by our method were not appearing in the lists of other known web search services, as well as many cases where the suggestions of commercial services like Google, Yahoo! and Bing were actually obsolete and rather outdated. In addition, our suggestions are based on common appearances of Twitter Entities (e.g. hashtags, mentions/replies to others

[34] https://musicbrainz.org/
[35] http://www.linkedmdb.org/
[36] http://wifo5-03.informatik.uni-mannheim.de/flickrwrappr/

users, web content via tinyURLs) in human annotated content (tweets). Such a content is rapidly disseminated, while if it is of great interest, it is maintained and reproduced many times, reused with other TEs, thus forming a dynamic network of resilient content capable of creating trends, top news, thematic categorizations, top-influencers, etc.. We demonstrated in this work how the most viral part of this network can be used in order to suggest related terms with respect to a user query. Finally, we ended up with some quite promising evaluations. The first one enrolled human raters and subjective comparisons of suggested results, with respect to related terms provided by Google for similar news events. The second evaluation provided us comparative results that clearly show that our work does not exactly belong solely to the generic query-suggestion research category, since it provides a broader, semantic-based view on it and most importantly it additionally utilizes the popularity/virality of Twitter Entities in the process. The high-level TEs that we aim to identify and suggest are characterized as "semantic entities carrying meaningful information", rather than "meaningless pieces of information".

Thus, the main contributions towards the construction of the proposed methodology can be identified with respect to five different aspects:

a. is based on a novel methodology, since it introduces the benefits of viral social content for the query suggestion problem,
b. is based on a solid research foundation, since it was evaluated against many famous search services, and user subjective ratings, as well as against similar baseline approaches and traditional information retrieval metrics,
c. successfully incorporates multiple types of information knowledge (e.g., temporal, textual and social content),
d. advances typical query suggestion methodologies by taking into account the concept of collective intelligence, as well as
e. further exploits the notion of Twitter Entities with respect to the query analysis research task.

In the future we aim to further investigate the benefits of our proposal, as a resource / social suggestion mechanism. Towards the first goal, we are working on a service where given a query (seed) the system will recommend possible URIs and multimedia content from well-known and popular social networks, such as ImgUr[37], YouTube[38] and others. With respect to social suggestion, we are working in a similar way to suggest high-influencers through the social sphere (e.g. follow a user, account), or to provide recommendations for joining a group and/or community. In parallel to the above, we plan to compare the provided query suggestions with other approaches that consider time as a virality factor when generating suggestion terms along with the recency factor evaluated in this work.

In addition, we intent to further investigate the feasibility of defining an adaptive mechanism capable of selecting the top-*k%* viral Twitter Entities when forming the Query Selection Set, given how viral a user's query is. Also it will be very interesting to investigate the impact of this value (k), as a trade-off parameter between the

[37] http://imgur.com/
[38] http://www.youtube.com/

system's latency (towards fast query provision) and classical information retrieval metrics (e.g. Mean Average Precision). Also, given the fact that query suggestion resembles the cluster labeling problem, our work can be used for enhancing unstructured microblogging messages similarly to the work of (Ounis et al., 2012) and others that deal with same issues. Finally, our future work includes issues such as an extension towards the semantification of the query suggestion mechanism. This practically means the association of related or synonymous hashtags for future queries, the hierarchical expression of types and relationships among suggested terms, as well as the involvement of well-known semantic vocabularies (e.g. FOAF[14], Dublin Core Metadata Initiative (DCMI)[39]).

[39] http://dublincore.org/documents/dcmi-terms/

Bibliography

(Abel et al., 2011)	Abel, F, Gao, Q, Houben, GJ and Tao, K 2011, 'Semantic enrichment of twitter posts for user profile construction on the social web', in G Antoniou et al. (eds), *Proceedings of the 8th extended semantic web conference on The semanic web: research and applications (ESWC, '11)*, Springer-Verlag, Berlin, Heidelberg, pp. 375-389. https://doi.org/10.1007/978-3-642-21064-8_26
(AlFalahi et al., 2013)	AlFalahi, K, Atif, Y and Harous, S 2013, 'Community detection in social networks through similarity virtual networks', *in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13)*, ACM, New York, NY, USA pp. 1116-1123. https://doi.org/10.1145/2492517.2500299
(Al-garadi et al., 2016)	Al-garadi, MA, Varathan, KD, Ravana, SD, Ahmed, E and Chang, V, 2016 'Identifying the influential spreaders in multilayer interactions of online social networks', *Journal of Intelligent and Fuzzy Systems*, vol. 31, no. 5, pp. 2721-2735. http://dx.doi.org/10.3233/JIFS-169112
(Al-garadi et al., 2017)	Al-garadi, MA, Varathan, KD and Ravana, SD2017, 'Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method', *Physica A: Statistical Mechanics and its Applications*, vol. 468, pp. 278-288. http://dx.doi.org/10.1016/j.physa.2016.11.002
(Almgren and Lee, 2016)	Almgren, K and Lee, J 2016, 'Applying an Influence Measurement Framework to Large Social Network', *Journal of Networking Technology*, vol. 7, pp. 6-15.
(Anagnostopoulos et al., 2008)	Anagnostopoulos, A, Kumar, R and Mahdian, M 2008, 'Influence and correlation in social networks', *in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*, ACM, New York, NY, USA, pp. 7-15. https://doi.org/10.1145/1401890.1401897
(Anagnostopoulos et al., 2012)	Anagnostopoulos I, Kolias, V and Mylonas, P 2012, 'Socio-semantic query expansion using Twitter hashtags', *in Proceedings of the 2012 7th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP '12)*, Luxembourg, pp. 29-34. https://doi.org/10.1109/SMAP.2012.15
(Anagnostopoulos et al., 2013)	Anagnostopoulos, I, Razis, G, Mylonas, P and Anagnostopoulos, CN 2013, 'Query Expansion with a Little Help from Twitter', *Engineering Applications of Neural Networks, Communications in Computer and Information Science*, vol. 384, pp. 20-29. http://dx.doi.org/10.1007/978-3-642-41016-1_3

(Anagnostopoulos et al., 2015)	Anagnostopoulos, I, Razis, G, Mylonas, P and Anagnostopoulos, CN 2015, 'Semantic query suggestion using Twitter Entities', *Neurocomputing*, vol. 163, pp. 137-150. https://doi.org/10.1016/j.neucom.2014.12.090
(Anger and Kittl, 2011)	Anger, I and Kittl, C 2011, 'Measuring influence on Twitter', in S Lindstaedt and M Granitzer (Eds), *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW '11)*, ACM, New York, NY, USA, pp. 1-4. http://dx.doi.org/10.1145/2024288.2024326
(Azaza et al., 2016)	Azaza, L, Kirgizov S, Savonnet, M, Leclercq, É, Gastineau, N and Rim, F 2016, 'Information fusion-based approach for studying influence on Twitter using belief theory', *Computational Social Networks*, vol. 3, no. 5. http://dx.doi.org/10.1186/s40649-016-0030-2
(Bai et al., 2015)	Bai, G, Liu L, Sun, B and Fang, J 2015, 'A survey of user classification in social networks', *in 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, pp. 1038-1041. http://dx.doi.org/10.1109/ICSESS.2015.7339230
(Bakshy et al., 2011)	Bakshy, E, Hofman, JM, Mason, WA and Watts, DJ 2011, 'Everyone's an influencer: quantifying influence on twitter', *in Proceedings of the fourth ACM international conference on Web search and data mining (WSDM, '11)*, ACM, New York, NY, USA, pp. 65-74. https://doi.org/10.1145/1935826.1935845
(Bakshy et al., 2012)	Bakshy, E, Rosenn, I, Marlow, C and Adamic, LA 2012, 'The role of social networks in information diffusion', *in Proceedings of the 21st international conference on World Wide Web (WWW, '12)*, ACM, New York, NY, USA, pp. 519-528. http://dx.doi.org/10.1145/2187836.2187907
(Bandyopadhyay et al., 2012)	Bandyopadhyay, A, Ghosh, K, Majumder, P and Mitra, M 2012, 'Query expansion for microblog retrieval', *International Journal of Web Science (IJWS)*, vol. 1, no. 4, pp. 368-380. http://dx.doi.org/10.1504/IJWS.2012.052535
(Barbieri et al., 2013)	Barbieri, N, Bonchi, F and Manco, G 2013, 'Cascade-based community detection', *in Proceedings of the sixth ACM international conference on Web search anddata mining (WSDM, '13)*, ACM, New York, NY, USA, pp. 33-42. http://dx.doi.org/10.1145/2433396.2433403
(Bauer and Kaltenböck, 2012)	Bauer, F and Kaltenböck, M 2012, 'Linked Open Data: The Essentials', Vienna, Austria, edition mono/monochrom, ISBN: 978-3-902796-05-9.
(Beiming and Ng, 2012)	Beiming, S and Ng, VTY 2012, 'Identifying Influential Users by their Postings in Social Networks', in M Atzmueller et al. (eds), *Ubiquitous Social Media Analysi,. MUSE 2012, MSM 2012*, Lecture Notes in Computer Science, vol. 8329, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-45392-

2_7

(Besel et al., 2016) Besel, C, Schlötterer, J and Granitzer, M 2016, 'On the quality of semantic interest profiles for online social network consumers', *SIGAPP Applied Computing Review*, vol. 16 no. 3, pp. 5-14. https://doi.org/10.1145/3015297. 3015298

(Berners-Lee, 2012) Tim BL 2012, '5 Star Open Data', http://5stardata.info/ (last access: October 2016)

(Biancalana et al., 2013) Biancalana, C, Gasparetti, F, Micarelli, A and Sansonetti, G 2013, 'Social semantic query expansion', *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, article 60, pp. 1-43. http://dx.doi.org/10.1145/2508037.2508041

(Bigonha et al., 2012) Bigonha, C, Cardoso, TNC, Moro, MM, Gonçalves, MA and Virgílio, AF 2012, 'Sentiment-based influence detection on Twitter', *Journal of the Brazilian Computer Society*, vol. 18, no. 3, pp. 169-183. https://doi.org/10.1007/s13173-011-0051-5

(Blei et al., 2003) Blei, D, Ng AY and Jordan, MI 2003, 'Latent dirichlet allocation', *The Journal of Machine Learning Research*, vol. 3, pp. 993-1022.

(Bouadjenek et al., 2016) Bouadjenek, MR, Hacid H and Bouzeghoub, M 2016, 'Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms', *Information Systems*, vol. 56, pp. 1-18. http://dx.doi.org/10.1016/j.is.2015.07.008

(Bouguessa and Romdhane, 2015) Bouguessa, M and Romdhane, LB 2015, 'Identifying authorities in online communities', *ACM Transactions on Intelligent Systems and Technology*, vol. 6 no. 3, article 30, pages 23. http://dx.doi.org/10.1145/2700481

(Boyd et al., 2010) Boyd, D, Golder, S and Lotan, G 2010, 'Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter', *in Proceedings of the 2010 43rd Hawaii International Conference on System Sciences (HICSS, '10)*, Honolulu, HI, pp. 1-10. http://dx.doi.org/10.1109/HICSS.2010.412

(Brickley and Miller, 2004) Brickley, D and Miller, L 2004, 'FOAF Vocabulary Specification. Namespace Document', FOAF Project, viewed 21 March 2018, <http://xmlns.com/foaf/0.1/>.

(Carvalho et al., 2017) Carvalho, JP, Rosa, H, Brogueira, G and Batista, F 2017, 'MISNIS: An intelligent platform for twitter topic mining', *Expert Systems with Applications*, vol. 89, pp 374-388. http://dx.doi.org/10.1016/j.eswa.2017.08.001

(Celik et al., 2011) Celik, I, Abel, F and Houben, GJ 2011, 'Learning Semantic Relationships between Entities in Twitter', in S. Auer et al. (eds), *Web Engineering, ICWE 2011*, Lecture Notes in Computer Science, Springer, vol. 6757, pp. 167-181.

(Cha et al., 2010) Cha, M, Haddadi, H, Benevenuto, F and Gummadi, KP 2010, 'Measuring User Influence in Twitter: The Million Follower

Fallacy', *in Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM 2010)*, Washington, DC, USA, May 23-26

(Chai et al., 2013) Chai, W, Xu, W, Zuo, M and Wen, X 2013, 'ACQR: A novel framework to identify and predict influential users in micro-blogging', in J Lee et al. (eds.), *17th pacific Asia conference on information systems, PACIS 2013*, Korea, pages 20. http://aisel.aisnet.org/ pacis2013/20

(Chen et al., 2010) Chen, J, Nairn, R, Nelson, L, Bernstein, M and Chi, E 2010, 'Short and tweet: experiments on recommending content from information streams', *in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI, '10)*, ACM, New York, NY, USA, pp. 1185-1194. https://doi.org/10.1145/1753326.1753503

(Chen et al., 2012) Chen, K, Chen, T, Zheng, G, Jin O, Yao, E and Yu, Y 2012, 'Collaborative personalized tweet recommendation', *in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR, '12)*, ACM, New York, NY, USA, pp. 661-670. https://doi.org/10.1145/2348283.2348372

(Chen et al., 2016a) Chen, H, Cui, X and Jin, H 2016, 'Top-k followee recommendation over microblogging systems by exploiting diverse information sources', *Future Generation Computer Systems*, vol. 55, no. C, pp. 534-543. https://doi.org/10.1016/j.future.2014.05.002

(Chen et al., 2016b) Chen, CC, Shih, SY and Lee, M 2016, 'Who should you follow? Combining learning to rank with social influence for informative friend recommendation', *Decision Support Systems*, vol. 90, pp. 33-45. http://dx.doi.org/10.1016/j.dss.2016.06.017

(Cheng et al., 2010) Cheng, Z, Caverlee, J and Lee, K 2010, 'You are Where you Tweet: a Content-based Approach to Geo-locating Twitter Users', *in Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM, '10)*, ACM, New York, NY, USA, pp.759-768. http://dx.doi.org/10.1145/ 1871437.1871535

(Chorley et al., 2015) Chorley, MJ, Colombo, GB, Allen, SM and Whitaker, RM 2015, 'Human content filtering in Twitter: The influence of metadata', *International Journal of Human-Computer Studies*, vol. 74, pp. 32-40. http://dx.doi.org/10.1016/j.ijhcs.2014.10.001

(Deb et al., 2016) Deb, B, Mukherjee, I, Srirama, SN and Vainikko, E 2016, 'A semantic followee recommender in Twitter using Topicmodel and Kalman filter', *in 12th IEEE International Conference on Control and Automation (ICCA)*, Kathmandu, pp. 649-656. http://dx.doi.org/10.1109/ ICCA.2016.7505352

(Dennett et al., 2016) Dennett, A, Nepal, S, Paris, C and Robinson B 2016, 'TweetRipple: Understanding Your Twitter Audience and the

Impact of Your Tweets', *in 2nd IEEE International Conference on Collaboration and Internet Computing (CIC)*, Pittsburgh, PA, USA, pp. 256-265. http://dx.doi.org/10.1109/CIC.2016.043

(Efron, 2010) Efron M 2010, 'Hashtag retrieval in a microblogging environment', *in Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR, '10)*, ACM, New York, NY, USA, pp. 787-788. http://dx.doi.org/10.1145/1835449.1835616

(Erlandsson et al., 2016) Erlandsson, F, Bródka, P, Borg, A and Johnson, H 2016, 'Finding Influential Users in Social Media Using Association Rule Learning', *Entropy*, vol. 18, pp. 164-179. http://dx.doi.org/10.3390/e18050164

(Fang and Hu, 2016) Fang, X and Hu, PJ 2016, 'Top Persuader Prediction for Social Networks', MIS Quarterly, Forthcoming, viewed May 3 2016, <https://ssrn.com/abstract=2774590>

(Francalanci and Husain, 2017) Francalanci, C and Hussain, A 2017, 'Influence-based Twitter browsing with NavigTweet', *Information Systems*, vol. 64, pp. 119-131. https://doi.org/10.1016/j.is.2016.07.012

(Gayo-Avello, 2013) Gayo-Avello, D 2013, 'Nepotistic relationships in Twitter and their impact on rank prestige algorithms', *Information Processing and Management*, vol. 49, no. 6, pp. 1250-1280. https://doi.org/10.1016/j.ipm.2013.06.003

(Geertzen, 2012) Geertzen, J 2012, 'Inter-rater agreement with multiple raters and variables', viewed August 5, 2018. https://mlnl.net/jg/software/ira/

(Goyal et al., 2010) Goyal, A, Bonchi, F and Lakshmanan, LVS 2010, 'Learning influence probabilities in social networks', *in Proceedings of the third ACM international conference on Web search and data mining (WSDM, '10)*, ACM, New York, NY, USA, pp. 241-250. http://dx.doi.org/10.1145/1718487.1718518

(Hajian et al., 2011) Hajian, B and White, T 2011, 'Modeling Influence in a Social Network: Metrics and Evaluation', *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, Boston, MA, pp. 497-500, http://dx.doi.org/10.1109/PASSAT/SocialCom.2011.118

(Hannon et al., 2010) Hannon, J, Bennett, M and Smyth, B 2010, 'Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches', *in Proceedings of the fourth ACM conference on Recommender systems (RecSys '10)*, ACM, New York, NY, USA, pp. 199-206. https://doi.org/10.1145/1864708.1864746

(Hassan et al., 2016) Hassan, Noura, El-Sharkawi, ME and El-Tazi, N 2016, 'Measuring User's Susceptibility to Influence in Twitter', *in Social Data Analytics and Management Workshop*, co-located with VLDB 2016, India

(Haralabopoulos et al., 2015)	Haralabopoulos, G, Anagnostopoulos and I, Zeadally, S 2015, 'Lifespan and propagation of information in On-line Social Networks: A case study based on Reddit', *Journal of Network and Computer* Applications, vol. 56, pp. 88-100. http://dx.doi.org/10.1016/j.jnca.2015.06.006
(Haralabopoulos et al., 2016)	Haralabopoulos, G, Anagnostopoulos, I and Zeadally, S 2016, 'The Challenge of Improving Credibility of User-Generated Content in Online Social Networks', *Journal of Data and Information Quality (JDIQ)*, vol. 7, no. 3, article 13, pp. 1-4. https://doi.org/10.1145/2899003
(Hepp, 2010)	Hepp, M 2010, 'HyperTwitter: Collaborative Knowledge Engineering via Twitter Messages', in P Cimiano and HS Pinto (eds), *Knowledge Engineering and Management by the Masses. EKAW 2010*, Lecture Notes in Computer Science, vol. 6317, Springer, pp. 451-461. http://dx.doi.org/10.1007/978-3-642-16438-5_35
(Hirsch, 2005)	Hirsch, JE 2005, 'An index to quantify an individual's scientific research output', *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, vol. 102, no. 46, USA, pp. 16569-16572. https://doi.org/10.1073/pnas.0507655102
(Hu et al., 2011)	Hu, X, Tang, L and Liu, H 2011, 'Enhancing accessibility of microblogging messages using semantic knowledge', in B Berendt et al. (eds), *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM, '11)*, ACM. NewYork. NY. USA. pp. 2465-2468. https://doi.org/10.1145/2063576.2063993
(Huang et al., 2013)	Huang, PY, Liu, HY, Lin, CT and Cheng, PJ 2013, 'A diversity-dependent measure for discovering influencers in social networks', in R E Banchs et al. (eds), *Information retrieval technology - 9th Asia information retrieval societies conference, AIRS 2013*, Singapore, December 9-11, in Lecture Notes in Computer Science, vol. 8281, Springer, pp. 368-379. https://doi.org/10.1007/978-3-642-45068-6_32
(Huang et al., 2016)	Huang, S, Zhang, J, Wang, L and Hua, XS 2016, 'Social Friend Recommendation Based on Multiple Network Correlation', *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 287-299. http://dx.doi.org/10.1109/ TMM.2015.2510333
(Huberman et al., 2009)	Huberman, BA, Romero, DM and Wu, F 2009, 'Social networks that matter: Twitter under the microscope', *First Monday*, vol. 14, no 1. https://doi.org/10.5210/fm.v14i1.2317
(Hung et al., 2016)	Hung, HJ, Yang, DN and Lee, WC 2016, 'Routing and Scheduling of Social Influence Diffusion in Online Social Networks', *in IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, Nara, pp. 437-446. http://dx.doi.org/10.1109/ICDCS.2016.45

(Jabeur et al., 2012) Jabeur, LB, Tamine, L and Boughanem, M 2012, 'Active microbloggers: Identifying influencers, leaders and discussers in microblogging networks', in L. Calderón-Benavides et al. (eds), *String processing and information retrieval - 19th international symposium, SPIRE 2012*, Cartagena de Indias, Colombia, October 21-25, Lecture Notes in Computer Science, Springer, vol. 7608, pp. 111-117. https://doi.org/10.1007/978-3-642-34109-0_12

(Jaitly et al., 2016) Jaitly, V, Chowriappa, P and Dua, S 2016, 'A framework to identify influencers in signed social networks', *in International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, pp. 2335-2340. http://dx.doi.org/10.1109/ICACCI.2016.7732403

(Kafeza et al., 2014) Kafeza, E, Kanavos, A, Makris, C and Vikatos, P 2014, 'Predicting Information Diffusion Patterns in Twitter', in L Iliadis et al. (eds), *Artificial Intelligence Applications and Innovations, AIAI 2014, IFIP Advances in Information and Communication Technology*, vol. 436, Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-662-44654-6_8

(Kafeza et al., 2016) Kafeza, E, Makris, C and Vikatos, P 2016, 'Marketing Campaigns in Twitter using a Pattern Based Diffusion policy', *in IEEE International Congress on Big Data (BigData Congress)*, San Francisco, CA, pp. 125-132. https://doi.org/10.1109/BigDataCongress.2016.24

(Kalloubi et al., 2016) Kalloubi, F, Nfaoui, EH and Elbeqqali, O 2016, 'Microblog semantic context retrieval system based on linked open data and graph-based theory', *Expert Systems With Applications*, vol. 53, no. C, pp.138-148. http://dx.doi.org/10.1016/j.eswa.2016.01.020

(Kanungsukkasem and Leelanupab, 2016) Kanungsukkasem, N and Leelanupab, T 2016, 'Power of Crowdsourcing in Twitter to Find Similar/Related Users', *in 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Khon Kaen, pp. 1-6. http://dx.doi.org/10.1109/JCSSE.2016.7748852

(Kapanipathi et al., 2014) Kapanipathi, P, Jain, P, Venkatramanihitra, C and Sheth, A 2014, 'User Interests Identification on Twitter Using a Hierarchical Knowledge Base', in V Presutti et al. (eds), *The Semantic Web: Trends and Challenges, ESWC 2014*, Lecture Notes in Computer Science, vol. 8465, Springer. http://dx.doi.org/10.1007/978-3-319-07443-6_8

(Karidi, 2016) Karidi, D 2016, 'From User Graph to Topics Graph', *in IEEE 32nd International Conference on Data Engineering Workshops (ICDEW)*, Helsinki, pp. 121-123. http://dx.doi.org/10.1109/ICDEW.2016.7495629

(Kimura and Saito, 2006) Kimura, M and Saito, K 2006, 'Tractable Models for Information Diffusion in Social Networks', in J Fürnkranz et al. (eds), *Knowledge Discovery in Databases: PKDD 2006*,

Lecture Notes in Computer Science, vol. 4213, Springer, pp. 259-271. http://dx.doi.org/10.1007/11871637_27

(King et al., 2013) King, D, Ramirez-Cano, D, Greaves, F, Vlaev, I, Beales, S and Darzi, A 2013, 'Twitter and the health reforms in the English National Health Service', *Health Policy*, vol. 110, no. 2, pp. 291-297. http://dx.doi.org/10.1016/j.healthpol.2013.02.005

(Kong and Feng, 2011) Kong, S and Feng, L 2011, 'A Tweet-Centric Approach for Topic-Specific Author Ranking in Micro-Blog', in J Tang et al. (eds), *Advanced Data Mining and Applications, ADMA 2011*, Springer, Berlin, Heidelberg, pp. 138-151. http://dx.doi.org/10.1007/978-3-642-25853-4_11

(Kumar and Carterette, 2013) Kumar, N and Carterette, B 2013, 'Time-based feedback and query expansion for twitter search', in Pl Serdyukov et al. (eds), *Proceedings of the 35th European Conferenceon Advances in Information Retrieval (ECIR, '13)*. Springer-Verlag. Berlin. Heidelberg. pp.734-737. https://doi.org/10.1007/978-3-642-36973-5_72

(Kumar et al., 2016) Kumar, N, Guo, R, Aleali, A and Shakarian, P 2016, 'An Empirical Evaluation Of Social Influence Metrics', *in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016)*, San Francisco, CA, USA, pp. 1329-1336. https://doi.org/10.1109/ASONAM.2016.7752408

(Kwak et al., 2010) Kwak, H, Lee, C, Park, H and Moon, S 2010, 'What is Twitter, a social network or a news media?', *in Proceedings of the 19th international conference on World wide web (WWW, '10)*, ACM, New York, NY, USA, pp. 591-600. http://dx.doi.org/10.1145/1772690.1772751

(Lampos et al., 2014) Lampos, V, Aletras, N, Preotiuc-Pietro, DT and Cohn, T 2014, 'Predicting and Characterising User Impact on Twitter', *in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL '14)*, pp. 405-413.

(Lavrenko and Croft, 2001) Lavrenko, V and Croft, WB 2001, 'Relevance based language models', *in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 120-127. https://doi.org/10.1145/ 383952.383972

(Lee and et al., 2010) Lee, C, Kwak, H, Park, H and Moon, S 2010, 'Finding influential based on the temporal order of information adoption in twitter', *in Proceedings of the WWW 2010*, Raleigh. NC, USA.

(Lehmann et al., 2015) Lehmann, J, Isele, R, Jakob, M, Jentzsch, A, Kontokostas, D, Mendes, PN, Hellmann, S, Morsey, M, Kleef, P van, Auer, S and Bizer, C 2015, 'DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia', *Semantic Web*, vol. 6, pp.167-195. https://doi.org/10.3233/SW-140134

(Lerman and Ghosh, 2010)	Lerman K and Ghosh R 2010, 'Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks', *in Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM '10)*, Washington, DC, USA
(Li et al., 2013a)	Li, L, Chen, X and Xu, G 2013, 'Suggestions for fresh search queries by mining microblog topics', in Cao L. et al. (eds) Behavior and Social Computing. BSIC 2013, BSI 2013. Lecture Notes in Computer Science, vol. 8178. Springer. pp. 214-223. https://doi.org/10.1007/978-3-319-04048-6_19
(Li et al., 2013b)	Li, X, Cheng, S, Chen, W and Jiang, F 2013, 'Novel user influence measurement based on user interaction in microblog', *in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13)*, ACM, New York, NY, USA, pp. 615-619. http://dx.doi.org/10.1145/2492517.2492635
(Li et al., 2017)	Li, D, Luo, Z, Ding, Y, Tang, J, Sun, GGZ, Dai, X, Du, J, Zhang, J and Kong, S 2017, 'User-level microblogging recommendation incorporating social influence', *Journal of the Association for Information Science and Technology*, vol. 68 no. 3, pp. 553-568. https://doi.org/10.1002/asi.23681
(Lim and Datta, 2013)	Lim, KH and Datta, A 2013, 'Interest classification of twitter users using Wikipedia', ACM 2013, *in Proceedings of the 9th International Symposium on Open Collaboration (WikiSym '13)*, ACM, New York, NY, USA, article 22, 2 pages. http://dx.doi.org/10.1145/2491055.2491078
(Liontis and Pitoura, 2016)	Liontis, K and Pitoura, E 2013, 'Boosting Nodes for Improving the Spread of Influence', https://arxiv.org/abs/1609.03478
(Liu et al., 2014)	Liu, N, Li, L, Xu, G and Yang, Z 2014, 'Identifying domain-dependent influential microblog users: A post-feature based approach', *in Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI '14)*, July 27-31, Québec city, Québec, Canada, pp. 3122-3123.
(Lu et al., 2016)	Lu, WX, Zhou, C and Wu, J 2016, 'Big social network influence maximization via recursively estimating influence spread', *Knowledge-Based Systems*, vol. 113, pp. 143-154. https://doi.org/10.1016/j.knosys.2016.09.020
(Ma et al., 2011)	Ma, H, Zhou, D, Liu, C, Lyu, MR and King, I 2011, 'Recommender systems with social regularization', *in Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 287-296.
(Massoudi et al., 2011)	Massoudi, K, Tsagkias, M, Rijke, M de and Weerkamp, W 2011, 'Incorporating query expansion and quality indicators in searching microblog posts', in P Clough et al. (eds), *Advances in Information Retrieval. ECIR 2011*, Lecture Notes in Computer Science, vol. 6611, pp. 362-367, Springer, Berlin,

Heidelberg. https://doi.org/10.1007/978-3-642-20161-5_36

(Miao et al., 2016) Miao, Q, Meng, Y and Sun, J 2016, 'Identifying the Most Influential Topic-Sensitive Opinion Leaders in Online Review Communities', *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu, pp. 330-335. http://doi.org/10.1109/ICCCBDA. 2016.7529579

(Michelson and Macskassy, 2010) Michelson, M and Macskassy, SA 2010, '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)*, ACM, New York, NY, USA, pp. 73-80. http://dx.doi.org/10.1145/1871840.1871852

(Mishne et al., 2013) Mishne, G, Dalton, J, Li, Z, Sharma, A and Lin, J 2013, 'Fast data in the era of big data: Twitter's real-time related query suggestion architecture', *in Proceedings of the 2013ACM SIGMOD International Conferenceon Management of Data (SIGMOD '13)*, ACM, New York, NY, USA, pp. 1147-1158. http://dx.doi.org/10.1145/2463676.2465290

(Mueller and Stumme, 2017) Mueller, J and Stumme, G 2017, 'Predicting Rising Follower Counts on Twitter Using Profile Information', *in 9th International ACM Web Science Conference 2017 (WebSci 2017)*, ACM, New York, NY, USA, pp. 121-130. https://doi.org/10.1145/3091478.3091490

(Nargundkar and Rao, 2016) Nargundkar, A and Rao, YS 2016, 'InfluenceRank A machine learning approach to measure influence of Twitter users', *in 2016 International Conference on Recent Trends in Information Technology (ICRTIT)*, Chennai, pp. 1-6. http://dx.doi.org/10.1109/ICRTIT.2016.7569535

(Naveed et al., 2011a) Naveed, N, Gottron, T, Kunegis, J and Alhadi, A 2011, 'Bad news travel fast: a content-based analysis of interestingness on Twitter', *in Proceedings of the 3rd International Web Science Conference (WebSci '11)*, ACM, New York, NY, USA, pp 8-15. http://dx.doi.org/10.1145/2527031.2527052

(Naveed et al., 2011b) Naveed, N, Gottron, T, Kunegis, J and Alhadi, AC 2011, 'Searching microblogs: coping with sparsity and document quality', *in Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ACM, pp.183-188. https://doi.org/10.1145/ 2063576.2063607

(Nigam et al., 2016) Nigam, A, Aguinaga, S and Chawla, NV 2016, 'Connecting the Dots to Infer Followers' Topical Interest on Twitter', *in International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, Durham, NC, pp. 1-6. https://doi.org/10.1109/BESC.2016.7804498

(Noia et al., 2012) Noia, TD, Mirizzi, R, Ostuni, VC, Romito, D and Zanker, M 2012, 'Linked open data to support content-based recommender systems', *I-SEMANTICS*, Graz, Austria, September 5-7, pp.1-8.

(Ohsaka et al., Ohsaka, N, Akiba, T, Yoshida, Y and Kawarabayashi, K 2016,

2016)	'Dynamic influence analysis in evolving networks', *Proceedings of the VLDB Endowment*, vol. 9, no. 12, pp. 1077-1088. http://dx.doi.org/10.14778/ 2994509.2994525
(Ounis et al., 2012)	Ounis, I, Macdonald, C, Lin, J and Soboroff, I 2012, 'Overview of the TREC 2011 Microblog Track', in *TREC*, <https://trec.nist.gov/pubs/trec20/papers/MICROBLOG.OVERVIEW.pdf>
(Overbey et al., 2013)	Overbey, LA, Paribello, C and Jackson, T 2013, 'Identifying Influential Twitter Users in the 2011 Egyptian Revolution', in AM Greenberg et al. (eds), *Proceedings of the 6th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP '13)*, Springer-Verlag, Berlin, Heidelberg, pp. 377-385. http://dx.doi.org/10.1007/978-3-642-37210-0_41
(Packer et al., 2012)	Packer, HS, Samangooei, S, Hare, J. S, Gibbins, N and Lewis, PH 2012, 'Event Detection using Twitter and Structured Semantic Query Expansion', *in Proceedings of the 1st international workshop on Multimodal crowd sensing (CrowdSens '12),* ACM, New York, NY, USA, pp. 7-1. http://dx.doi.org/10.1145/2390034.2390039
(Pal and Counts, 2011)	Pal, A and Counts, S 2011, 'Identifying topical authorities in microblogs', *in Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11)*, ACM, New York, NY, USA, pp. 45-54. https://doi.org/10.1145/1935826.1935843
(Passant, 2010)	Passant, A 2010, 'dbrec - Music Recommendations Using DBpedia', *The Semantic Web - (ISWC) 2010 - 9th International Semantic Web Conference*, Shanghai, China, November 7-11, pp. 209-224.
(Peng et al., 2017)	Peng, S, Yang, A, Cao, L, Yu, S and Xie, D 2017, 'Social influence modeling using information theory in mobile social networks', *Information Sciences*, vol. 379, pp. 146-159. http://dx.doi.org/10.1016/j.ins.2016.08.023
(Phelan et al., 2011)	Phelan, O, McCarthy, K, Bennett, M and Smyth, B 2011, 'Terms of a Feather: Content-Based News Recommendation and Discovery Using Twitter', in P Clough et al. (eds), *Advances in Information Retrieval, ECIR 2011*, Lecture Notes in Computer Science, vol. 6611, Springer, pp.448-459. http://dx.doi.org/10.1007/978-3-642-20161-5_44
(Piškorec et al., 2016)	Piškorec, M, Antulov-Fantulin, N, Miholic, I, Šmuc, T and Šikic, M 2016, 'Modeling peer and external influence in online social networks. https://arxiv.org/abs/1610.08262
(Poghosyan and Ifrim, 2016)	Poghosyan, G and Ifrim, G 2016, 'Real-time News Story Detection and Tracking with Hashtags', *in 2nd Workshop on Computing News Storylines*, Austin, Texas, USA, November 5, pp. 20-29.
(Pollock et al.,	Pollock, KH, Nichols, JD, Brownie C and Hines JE 1990,

1990) 'Statistical inference for capture-recapture experiments', *Wildlife Monographs 107*, pp. 3-97.

(Qi et al., 2012) Qi, GJ, Aggarwal, CC and Huang, T 2012, 'Community Detection with Edge Content in Social Media Networks', *in Proceedings of the 2012 IEEE 28th International Conference on Data Engineering (ICDE '12)*, IEEE Computer Society, Washington, DC, USA, pp. 534-545. http://dx.doi.org/10.1109/ICDE.2012.77

(Räbiger and Spiliopoulou, 2015) Räbiger, S and Spiliopoulou, M 2015, 'A framework for validating the merit of properties that predict the influence of a twitter user', *Expert Systems with Applications*, vol. 42, no. 5, pp. 2824-2834. http://dx.doi.org/10.1016/ j.eswa.2014.11.006

(Ramírez-de-la-Rosa et al., 2014) Ramírez-de-la-Rosa, G, Villatoro-Tello, E, Jiménez-Salazar, H and Sánchez-Sánchez, C 2014, 'Towards automatic detection of user influence in twitter by means of stylistic and behavioral features', in AF Gelbukh et al. (eds), *Human-inspired computing and its applications - 13th Mexican international conference on artificial intelligence, MICAI 2014*, Tuxtla Gutierrez, Mexico, November 16-22, Lecture Notes in Computer Science, Springer, vol. 8856, pp. 245-256. https://doi.org/10.1007/978-3-319-13647-9_23

(Razis and Anagnostopoulos, 2014a) Razis, G and Anagnostopoulos, I 2014, 'InfluenceTracker: Rating the impact of a Twitter account', *Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology*, vol. 437, pp. 184-195. https://doi.org/10.1007/978-3-662-44722-2_20

(Razis and Anagnostopoulos, 2014b) Razis, G and Anagnostopoulos, I 2014, 'Semantifying Twitter: the InfluenceTracker Ontology', *in 9th International Workshop on Semantic and Social Media Adaptation and Personalization*, Corfu, pp. 98-103, http://dx.doi.org/10.1109/SMAP.2014.23

(Razis and Anagnostopoulos, 2016) Razis, G and Anagnostopoulos, I 2016, 'Discovering similar Twitter accounts using semantics', *Engineering Applications of Artificial Intelligence*, vol. 51, pp. 37-49. http://dx.doi.org/10.1016/j.engappai.2016.01.015

(Razis et al., 2015) Razis, G, Anagnostopoulos, I and Vafopoulos, M 2015, 'Semantic social analytics and Linked Open Data cloud', *in 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Trento, pp. 1-6. http://dx.doi.org/10.1109/SMAP.2015.7370091

(Razis et al., 2016) Razis, G, Anagnostopoulos and I, Saloun, P 2016, 'Thematic labeling of Twitter accounts using DBpedia properties', *in 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Thessaloniki, pp. 106-111. http://dx.doi.org/10.1109/ SMAP.2016.7753393

(Razis et al., 2018) Razis, G, Anagnostopoulos and I, Sherali, Z 2018, 'Modeling Influence With Semantics in Social Networks: a Survey', *ACM*

Computing Surveys (under review), draft version at https://arxiv.org/abs/1801.09961

(Reda et al., 2011) Reda, MR, Hacid, H, Bouzeghoub, M and Daigremont, J 2011, 'Personalized Social Query Expansion Using Social Bookmarking Systems', *in Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*, Beijing, China, pp.1113-1114. http://dx.doi.org/10.1145/2009916.2010075

(Riquelme and Cantergiani, 2016) Riquelme, F and Cantergiani, PG 2016, 'Measuring user influence on Twitter: A Survey', *Information Processing and Management*, vol. 52, no. 5, pp. 949-975. https://doi.org/10.1016/j.ipm.2016.04.003

(Romero et al., 2011a) Romero, DM, Asur, W, Galuba, S and Huberman, BA 2011, 'Influence and Passivity in Social Media', *in Proceedings of the 20th international conference companion on World Wide Web (WWW '11)*, ACM, New York, NY, USA, pp. 113-114. http://dx.doi.org/10.1145/1963192.1963250

(Romero et al., 2011b) Romero, DM, Meeder, B and Kleinberg, J 2011, 'Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter', *in Proceedings of the 20th international conference on World Wide Web (WWW '11)*, ACM, New York, NY, USA, pp. 695-704. http://dx.doi.org/10.1145/1963405.1963503

(Singer, 2016) Singer, Y 2016, 'Influence maximization through adaptive seeding', *ACM SIGecom Exchanges*, vol. 15, no. 1, pp. 32-59. http://dx.doi.org/10.1145/2994501.2994503

(Shabir and Clarke, 2009) Shabir, N and Clarke, C 2009, 'Using linked data as a basis for a learning resource recommendation system', in Proceedings of EC-TEL, '09.

(Shinavier, 2010) Shinavier, J 2010, 'Real-time #Semantic Web in <= 140chars', *in Proceedings of WWW 2010*, April 26-30, Raleigh Convention Center, Raleigh, NC, USA.

(Slabbekoorn et al., 2016) Slabbekoorn, K, Noro, T and Tokuda, T 2016, 'Ontology-Assisted Discovery of Hierarchical Topic Clusters on the Social Web', *Journal of Web Engineering*, vol. 15, pp. 361-396.

(Spina et al., 2012) Spina, D, Meij, E, Rijke, M de, Oghina, A, Bui, MT and Breuss, M 2012, 'Identifying entity aspects in microblog posts', *in Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*, ACM, New York, NY, USA, pp.1089-1090. https://doi.org/10.1145/ 2348283.2348483

(Srinivasan et al., 2014) Srinivasan, MS, Srinivasa, S and Thulasidasan, S 2014, 'A Comparative Study of two Models for Celebrity Identification on Twitter', *in Proceedings of the 20th International Conference on Management of Data (COMAD '14)*, Computer

Society of India. Hyderabad, India, pp.57–65.

(Song et al., 2017) Song, G, Li, Y, Chen, X, He, X and Tang, J 2017, 'Influential Node Tracking on Dynamic Social Network: An Interchange Greedy Approach', *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 2, pp. 359-37. http://dx.doi.org/10.1109/TKDE.2016.2620141

(Subbian et al., 2016) Subbian, K, Aggarwal, C and Srivastava, J 2016, 'Mining Influencers Using Information Flows in Social Streams', *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 3, article 26. http://doi.acm.org/10.1145/2815625

(Sun and Ng, 2012) Sun, B and Ng, VTY 2012, 'Identifying influential users by their postings in social networks', *in Proceedings of the 3rd international workshop on Modeling social media (MSM '12)*, ACM, New York, NY, USA, pp. 1-8. http://dx.doi.org/10.1145/2310057.2310059

(Sun et al., 2009) Sun, E, Rosenn, I, Marlow CA and Lento, TM 2009, 'Gesundheit! modeling contagion through facebook news feed', *in Proceedings of International AAAI Conference on Weblogs and Social Media*, San Jose, CA, US.

(Tang et al., 2016) Tang, X, Miao, Q, Yu, S and Quan, Y 2016, 'A Data-Based Approach to Discovering Multi-Topic Influential Leaders', *PLOS ONE*, vol. 11, no. 7. http://dx.doi.org/10.1371/journal.pone.0158855

(Tao et al., 2012) Tao, K, Abel, F, Hau, C and Houben, GJ 2012, 'Twinder: A Search Engine for Twitter Streams', *in Proceedings of the 12th International Conference on Web Engineering (ICWE 2012)*, Berlin, Germany, pp. 153-168.

(Veijalainen et al., 2015) Veijalainen, J, Semenov, A and Reinikainen, M 2015, 'User Influence and Follower Metrics in a Large Twitter Dataset', *in Proceedings of the 11th International Conference on Web Information Systems and Technologies (WEBIST-2015)*, Lisbon, Portugal, pp. 487-497. http://dx.doi.org/10.5220/0005410004870497

(Vocht et al., 2011) Vocht, L de, Softic, S, Ebner, M and Mühlburger, H, 'Semantically Driven Social Data Aggregation Interfaces for Research 2.0', in S Lindstaedt and M Granitzer (eds.), *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW '11)*, ACM, New York, NY, USA, 2011, pp. 43-52. http://dx.doi.org/10.1145/2024288.2024339

(Wang et al., 2016) Wang, Y, Liu, J, Huang, Y and Feng, X 2016, 'Using Hashtag Graph-Based Topic Model to Connect Semantically-Related Words Without Co-Occurrence in Microblogs', *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, pp. 1919-1933. http://dx.doi.org/10.1109/TKDE.2016.2531661

(Wei et al., 2016a) Wei, W, Cong, G, Miao, C, Zhu, F and Li, G 2016, 'Learning to

Find Topic Experts in Twitter via Different Relations', *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1764-1778. http://dx.doi.org/10.1109/TKDE.2016.2539166

(Wei et al., 2016b) Wei, J, Mengdi, G, Xiaoxi, W and Xianda, W 2016, 'A new evaluation algorithm for the influence of user in Social Network', *China Communications*, vol. 13, no. 2, pp. 200-206. http://dx.doi.org/10.1109/CC.2016.7405737

(Weng et al., 2010) Weng, J, Lim, E. P, Jiang, J and He, Q 2010, 'TwitterRank: finding topic-sensitive influential twitterers', *in Proceedings of the third ACM international conference on Web search and data mining (WSDM '10)*, ACM, New York, NY, USA, pp. 261-270. http://dx.doi.org/10.1145/1718487.1718520

(Xia and Bu, 2012) Xia, ZY and Bu, Z 2012, 'Community detection based on a semantic network', *Knowledge Based Systems*, vol. 26, pp. 30-39. https://doi.org/10.1016/j.knosys.2011.06.014

(Yang and Counts, 2010) Yang, J and Counts, S 2010, 'Predicting the Speed, Scale, and Range of Information Diffusion in Twitter', *in Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM '10)*, Washington, DC, USA.

(Yang and Leskovec, 2010) Yang, J and Leskovec, J 2010, 'Modeling Information Diffusion in Implicit Networks', *in Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM '10)*, IEEE Computer Society, Washington, DC, USA, pp. 599-608. http://dx.doi.org/10.1109/ICDM.2010.22

(Yang et al., 2013) Yang, J, McAuley, J and Leskovec, J 2013, 'Community Detection in Networks with Node Attributes', *in IEEE 13th International Conference on Data Mining*, Dallas, TX, pp. 1151-1156. http://dx.doi.org/10.1109/ICDM.2013.167

(Yang et al., 2016) Yang, Y, Chawla, NV, Lichtenwalter, RN and Dong Y 2016, 'Influence Activation Model: A New Perspective in Social Influence Analysis and Social Network Evolution', https://arxiv.org/abs/1605.08410

(Yang et al., 2017) Yang, Y, Wang, Z, Pei, J and Chen, E 2017, 'Tracking Influential Nodes in Dynamic Networks', *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 11, pp. 2615-2628. https://doi.org/10.1109/TKDE.2017.2734667

(Yi et al., 2016) Yi, X, Shen, X, Lu, W, Chan, TS and Chung, FL 2016, 'Persuasion driven influence analysis in online social networks', *in International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada, pp. 4451-4456. http://dx.doi.org/10.1109/IJCNN.2016.7727782

(Yu et al., 2016) Yu, M, Yang, W, Wang, W and Shen, GW 2016, 'Information influence measurement based on user quality and information attribute in microblogging', *in 8th IEEE International Conference on Communication Software and Networks*

(ICCSN), Beijing, pp. 603-608. http://dx.doi.org/10.1109/ICCSN.2016.7586594

(Yuan et al., 2015) Yuan, T, Cheng, J, Zhang, X, Liu, Q and Lu, H 2015, 'How friends affect user behaviors? An exploration of social relation analysis for recommendation', *Knowledge-Based Systems*, vol. 88, pp. 70-84. http://dx.doi.org/10.1016/j.knosys.2015.08.005

(Zangerle et al., 2015) Zangerle, E, Schmidhammer, G and Specht, G 2015, '#Wikipedia on Twitter analyzing tweets about Wikipedia', *in Proceedings of the 11th International Symposium on Open Collaboration, (OpenSym '15)*, ACM, New York, NY, USA, article 14, pp. 14:1-14:8. https://doi.org/10.1145/2788993.2789845

(Zhang et al., 2016) Zhang, Q, Wu, J, Yangx, H, Luy, W, Long, G and Zhang, C 2016, 'Global and Local Influence-based Social Recommendation', *in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*, Indianapolis, USA, October 24-28, pp. 1917-1920. http://dx.doi.org/10.1145/ 2983323.2983873

(Zhang et al., 2017) Zhang, Y, Moe, WW and Schweidel, DA 2017, 'Modeling the role of message content and influencers in social media rebroadcasting', *International Journal of Research in Marketing*, vol. 34, no. 1, pp. 100-119. http://dx.doi.org/10.1016/j.ijresmar.2016.07.003

(Zhao et al., 2014) Zhao, K, Yen, J, Greer, G, Qiu, B, Mitra, P and Portier, K 2014, 'Finding influential users of online health communities: a new metric based on sentiment influence', *Journal of the American Medical Informatics Association*, vol. 21, no e2, pp e212-e218. http://dx.doi.org/10.1136/amiajnl-2013-002282

(Zhou et al., 2012) Zhou, D, Lawless, S and Wade, V 2012, 'Improving search via personalized query expansion using social media', *Information Retrieval*, vol. 15, no. 3-4, pp. 218-242. https://doi.org/10.1007/s10791-012-9191-2

(Zhuang et al., 2017) Zhuang, K, Shen, H and Zhang, H 2017, 'User spread influence measurement in microblog', *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 3169-3185. http://dx.doi.org/10.1007/s11042-016-3818-z

Appendix A: Comparison of Reviewed Articles

Table 2.3: Classification of referenced works. Fs and Fing refer to follow-up relationships, P to posts, RP to reposts, FL to favorite or liked posts, M to mentions, R to replies, T to topics, and CA to content analysis.

Category	Section	Reference	Relations		Activities					Context	
			Fs	Fing	P	RP	FL	M	R	T	CA
Direct social information-based	2.4.1.1	(Anger and Kittl, 2011)	✓	×	✓	✓	×	✓	✓	×	×
		(King et al., 2013)	×	×	×	✓	×	×	×	×	×
		(Cha et al., 2010)	✓	×	✓	✓	×	✓	×	✓	×
		(Romero et al., 2011a)	✓	×	✓	✓	×	×	×	×	×
		(Hassan et al., 2016)	×	×	✓	✓	✓	✓	✓	✓	×
		(Peng et al., 2017)	×	×	×	×	×	×	✓	×	×
		(Razis and Anagnostopoulos, 2014a)	✓	✓	✓	✓	✓	×	×	×	×
		(Razis and Anagnostopoulos, 2014b)	✓	✓	✓	✓	✓	×	×	×	×
		(Dennett et al., 2016)	✓	×	✓	×	×	✓	✓	×	✓

Category	Section	Reference	Relations		Activities					Context	
			Fs	Fing	P	RP	FL	M	R	T	CA
Hyperlink-based	2.4.1.2	(Hajian et al., 2011)	✓	✓	×	✓	✓	✓	×	×	×
		(Li et al., 2013b)	✓	✓	✓	✓	×	✓	×	×	✓
		(Jabeur et al., 2012)	✓	×	×	✓	×	✓	×	×	×
		(Wei et al., 2016b)	✓	✓	✓	✓	×	×	✓	×	✓
		(Carvalho et al., 2017)	×	×	×	×	×	✓	×	✓	✓
Machine Learning techniques-based	2.4.1.3	(Lampos et al., 2014)	✓	✓	✓	×	×	✓	✓	×	×
		(Nargundkar and Rao, 2016)	✓	✓	✓	×	×	×	×	×	×
		(Fang and Hu, 2016)	×	×	×	×	×	×	×	×	×
		(Mueller and Stumme, 2017)	×	×	×	×	×	×	×	×	✓

Category	Section	Reference	Relations		Activities					Context	
			Fs	Fing	P	RP	FL	M	R	T	CA
Propagation-oriented Approaches	2.4.2.1	(Yuan et al., 2015)	✓	✓	✓	×	×	×	×	×	×
		(Huang et al., 2013)	✓	×	×	✓	×	×	×	×	×
		(Jabeur et al., 2012)	✓	×	×	✓	×	✓	×	×	×
		(Romero et al., 2011a)	✓	×	✓	✓	×	×	×	×	×
		(Barbieri et al., 2013)	✓	×	✓	✓	×	×	×	×	×
		(Yang and Counts, 2010)	×	×	×	✓	×	✓	✓	✓	×
		(Yang and Leskovec, 2010)	×	×	×	✓	×	×	×	×	×
		(Kimura and Saito, 2006)	✓	×	✓	✓	×	×	×	×	×
		(Bakshy et al., 2012)	✓	✓	×	✓	×	×	✓	×	×
		(Tang et al., 2016)	✓	✓	✓	✓	×	×	×	✓	✓
		(Jaitly et al., 2016)	✓	✓	×	×	×	×	×	×	×
		(Lu et al., 2016)	✓	×	✓	×	×	×	×	×	×
		(Song et al., 2017)	✓	×	×	×	×	×	×	×	×
		(Yang et al., 2016)	✓	×	×	✓	×	×	×	×	×
		(Subbian et al., 2016)	✓	✓	×	✓	×	×	×	×	✓
		(Yi et al., 2016)	✓	✓	×	×	×	×	×	✓	×
		(Hung et al., 2016)	✓	✓	×	✓	×	×	×	×	×
		(Liontis and Pitoura, 2016)	✓	×	×	×	×	×	×	×	×

Category	Section	Reference	Relations		Activities					Context	
			Fs	Fing	P	RP	FL	M	R	T	CA
Diffusion-oriented Approaches	2.4.2.2	(Bakshy et al., 2011)	✓	×	×	✓	×	×	×	×	×
		(Sun et al., 2009)	✓	✓	✓	×	×	×	×	×	×
		(Kwak et al., 2010)	✓	×	✓	✓	×	×	×	×	×
		(Romero et al., 2011a)	✓	×	✓	✓	×	×	×	×	×
		(Naveed et al., 2011a)	✓	×	×	✓	×	×	×	×	✓
		(Al-garadi et al., 2017)	✓	×	×	✓	×	✓	×	×	×
		(Yang and Counts, 2010)	×	×	×	✓	×	✓	✓	✓	×
		(Lerman and Ghosh, 2010)	✓	✓	✓	✓	×	×	×	×	×
		(Bakshy et al., 2012)	✓	✓	×	✓	×	×	✓	×	×
		(Romero et al., 2011b)	×	×	✓	×	×	✓	×	✓	×
		(Wei et al., 2016b)	✓	✓	✓	✓	×	×	✓	×	✓
		(Al-garadi et al., 2016)	✓	×	×	✓	×	✓	×	×	×
		(Veijalainen et al., 2015)	✓	×	×	✓	×	×	×	×	×
		(Zhuang et al., 2017)	✓	×	×	✓	×	×	✓	×	×
		(Kafeza et al., 2016)	×	×	✓	✓	×	×	×	×	×
		(Haralabopoulos et al., 2015)	×	×	✓	×	×	×	×	×	×
		(Kafeza et al., 2014)	×	×	✓	✓	×	×	×	×	×

Category	Section	Reference	Relations		Activities					Context	
			Fs	Fing	P	RP	FL	M	R	T	CA
Network / Graph Properties	2.4.3	(Overbey et al., 2013)	×	×	×	×	×	✓	✓	✓	×
		(Jabeur et al., 2012)	✓	×	×	✓	×	✓	×	×	×
		(AlFalahi et al., 2013)	✓	×	✓	×	×	×	×	×	✓
		(Karidi, 2016)	×	×	✓	×	×	×	×	×	✓
		(Kalloubi et al., 2016)	×	×	✓	×	×	×	×	×	✓
		(Jaitly et al., 2016)	✓	✓	×	×	×	×	×	×	×
		(Wei et al., 2016b)	✓	✓	✓	✓	×	×	✓	×	✓
		(Almgren and Lee, 2016)	✓	✓	✓	×	×	×	×	×	×
		(Ohsaka et al., 2016)	✓	✓	×	×	×	×	×	×	×
		(Song et al., 2017)	✓	×	×	×	×	×	×	×	×
		(Yang et al., 2016)	✓	×	×	✓	×	×	×	×	×
		(Yang et al., 2017)	✓	✓	×	×	×	×	×	×	×

Category	Section	Reference	Relations		Activities					Context	
			Fs	Fing	P	RP	FL	M	R	T	CA
Ranking	2.4.4.1	(Kong and Feng, 2011)	×	×	✓	✓	×	×	×	✓	×
		(Li et al., 2013b)	✓	✓	✓	✓	×	✓	×	×	✓
		(Wei et al., 2016a)	✓	×	✓	×	×	×	×	✓	×
		(Miao et al., 2016)	✓	×	✓	×	×	×	×	✓	✓
		(Francalanci and Husain, 2017)	✓	✓	✓	✓	✓	✓	×	×	×
		(Chen et al., 2016b)	✓	✓	✓	×	✓	×	✓	×	×
		(Zhuang et al., 2017)	×	×	×	✓	×	✓	✓	×	×
		(Razis and Anagnostopoulos, 2014a)	✓	✓	✓	✓	✓	×	×	×	×
		(Razis and Anagnostopoulos, 2014b)	✓	✓	✓	✓	✓	×	×	×	×

Category	Section	Reference	Relations		Activities					Context	
			Fs	Fing	P	RP	FL	M	R	T	CA
Recommendation	2.4.4.2	(Yuan et al., 2015)	✓	✓	✓	×	×	×	×	×	×
		(Phelan et al., 2011)	✓	✓	✓	×	×	×	×	×	✓
		(Hannon et al., 2010)	✓	✓	✓	×	×	×	×	×	✓
		(Chen et al., 2010)	×	✓	✓	×	×	×	×	✓	✓
		(Ma et al., 2011)	✓	✓	✓	×	×	×	×	×	×
		(Huang et al., 2016)	✓	✓	×	×	×	×	×	×	✓
		(Kanungsukkasem and Leelanupab, 2016)	×	×	×	×	×	×	×	×	×
		(Deb et al., 2016)	✓	✓	✓	×	×	×	×	×	✓
		(Chen et al., 2012)	✓	✓	✓	✓	×	✓	×	×	✓
		(Karidi, 2016)	×	×	✓	×	×	×	×	×	✓
		(Chen et al., 2016a)	✓	✓	✓	✓	×	×	×	×	✓
		(Tang et al., 2016)	✓	✓	✓	✓	×	×	×	✓	✓
		(Zhang et al., 2016)	✓	✓	✓	×	×	×	×	×	×
		(Chen et al., 2016b)	✓	✓	✓	×	✓	×	✓	×	×
		(Li et al., 2017)	✓	✓	×	×	×	×	×	✓	×
Other Application Domains	2.4.4.3	(Bigonha et al., 2012)	✓	✓	✓	✓	✓	✓	×	×	✓
		(Zhao et al., 2014)	×	×	✓	×	✓	×	×	✓	×
		(Poghosyan and Ifrim, 2016)	×	×	✓	×	×	×	×	×	✓
		(Piškorec et al., 2016)	✓	✓	×	×	×	×	×	×	×

Table 2.4: Classification of referenced works. NS refers to network structure, (R)P to (re)-posts, I to interactions, P to profiling and personalization, T to topics, O to ontologies, and H to hashtags.

Category	Section	Reference	NS	Activities		Context		KB	Semantic Web		Entities	
				(R)P	I	P	T		RDF	O	H	URLs
Social Modeling	2.5.1	(Celik et al., 2011)	×	✓	×	×	×	✓	×	×	×	×
		(Abel et al., 2011)	×	✓	×	✓	×	×	✓	×	✓	✓
		(Shinavier, 2010)	×	✓	×	×	×	×	✓	✓	✓	✓
		(Vocht et al., 2011)	×	✓	×	×	✓	✓	✓	✓	×	×
		(Xia and Bu, 2012)	✓	✓	✓	×	×	×	×	×	×	×
		(Packer et al., 2012)	×	✓	×	×	✓	✓	✓	×	×	×
		(Hepp, 2010)	×	✓	×	×	×	×	✓	×	✓	×
		(Slabbekoorn et al., 2016)	✓	✓	×	×	✓	✓	×	✓	×	×
		(Deb et al., 2016)	✓	✓	×	✓	×	✓	×	×	×	×
		(Karidi, 2016)	×	✓	×	✓	×	✓	×	×	×	×
		(Besel et al., 2016)	✓	✓	×	✓	×	✓	×	×	×	×
		(Kalloubi et al., 2016)	✓	✓	×	×	✓	✓	×	×	×	×
		(Wang et al., 2016)	×	✓	×	×	✓	×	×	×	✓	×
		(Razis and Anagnostopoulos, 2014b)	✓	✓	✓	×	×	×	✓	✓	✓	✓
		(Razis et al., 2015)	✓	✓	✓	×	×	✓	✓	✓	✓	✓
		(Razis et al., 2016)	×	×	×	✓	×	✓	×	×	×	×

Category	Section	Reference	NS	Activities		Context		KB	Semantic Web		Entities	
				(R)P	I	P	T		RDF	O	H	URLs
User-oriented	2.5.2.1	(Räbiger and Spiliopoulou, 2015)	✓	✓	✓	×	×	×	×	×	×	×
		(Sun and Ng, 2012)	✓	✓	×	×	✓	×	×	×	×	×
		(Kong and Feng, 2011)	✓	✓	×	×	✓	×	×	×	×	×
		(Overbey et al., 2013)	✓	×	✓	×	✓	×	×	×	×	×
		(Weng et al., 2010)	✓	×	×	×	✓	×	×	×	×	×
		(Liu et al., 2014)	✓	✓	×	×	✓	×	×	×	×	×
		(Ramírez-de-la Rosa et al., 2014)	×	✓	✓	×	×	×	×	×	✓	✓
		(Celik et al., 2011)	×	✓	×	×	×	✓	×	×	×	×
		(Abel et al., 2011)	×	✓	×	✓	×	×	✓	×	✓	✓
		(Vocht et al., 2011)	×	✓	×	×	✓	✓	✓	✓	×	×
		(Ma et al., 2011)	✓	✓	×	×	×	×	×	×	×	×
		(Packer et al., 2012)	×	✓	×	×	✓	✓	✓	×	×	×
		(Zhou et al., 2012)	×	✓	×	✓	×	×	×	×	✓	×
		(Redaet et al., 2011)	×	×	×	✓	×	×	×	×	×	×
		(Efron, 2010)	×	✓	×	×	×	×	×	×	✓	×
		(Slabbekoorn et al., 2016)	✓	✓	×	×	✓	✓	×	✓	×	×
		(Huang et al., 2016)	✓	×	×	×	✓	×	×	×	×	×

Category	Section	Reference	NS	Activities		Context		KB	Semantic Web		Entities	
				(R)P	I	P	T		RDF	O	H	URLs
User-oriented (cont.)	2.5.2.1	(Kanungsukkasem and Leelanupab, 2016)	✓	×	×	×	×	×	×	×	×	×
		(Nigam et al., 2016)	×	✓	×	✓	×	×	×	×	×	✓
		(Karidi, 2016)	×	✓	×	✓	×	✓	×	×	×	×
		(Besel et al., 2016)	✓	✓	×	✓	×	✓	×	×	×	×
		(Poghosyan and Ifrim, 2016)	×	✓	×	×	×	×	×	×	✓	×
		(Miao et al., 2016)	✓	✓	×	✓	×	×	×	×	×	×
User-oriented	4.2.1	(Hassan et al., 2016)	×	✓	✓	×	×	×	×	×	×	×
		(Zhang et al., 2017)	×	✓	×	✓	×	×	×	×	×	×
		(Fang and Hu, 2016)	✓	×	✓	×	×	×	×	×	×	×
		(Razis and Anagnostopoulos, 2016)	×	✓	✓	×	×	×	✓	✓	✓	✓
		(Anagnostopoulos et al., 2015)	×	✓	✓	×	×	×	×	×	✓	✓
		(Anagnostopoulos et al., 2013)	×	✓	✓	×	×	×	×	×	✓	✓
		(Carvalho et al., 2017)	✓	✓	✓	×	✓	×	×	×	✓	×

Category	Section	Reference	NS	Activities		Context		KB	Semantic Web		Entities	
				(R)P	I	P	T		RDF	O	H	URLs
Topic and Event-oriented	2.5.2.2	(Räbiger and Spiliopoulou, 2015)	✓	✓	✓	×	×	×	×	×	×	×
		(Sun and Ng, 2012)	✓	✓	×	×	✓	×	×	×	×	×
		(Kong and Feng, 2011)	✓	✓	×	×	✓	×	×	×	×	×
		(Pal and Counts, 2011)	✓	✓	✓	×	✓	×	×	×	✓	✓
		(Michelson and Macskassy, 2010)	×	✓	✓	✓	✓	✓	×	×	×	×
		(Lim and Datta, 2013)	✓	✓	×	✓	×	✓	×	×	✓	×
		(Kapanipathi et al., 2014)	×	✓	×	✓	×	✓	×	×	×	×
		(Tao et al., 2012)	✓	✓	✓	×	✓	✓	×	×	✓	✓
		(Packer et al., 2012)	×	✓	×	×	✓	✓	✓	×	×	×
		(Slabbekoorn et al., 2016)	✓	✓	×	×	✓	✓	×	✓	×	×
		(Deb et al., 2016)	✓	✓	×	✓	×	✓	×	×	×	×
		(Nigam et al., 2016)	×	✓	×	✓	×	×	×	×	×	✓
		(Karidi, 2016)	×	✓	×	✓	×	✓	×	×	×	×
		(Wei et al., 2016a)	✓	✓	×	×	✓	×	×	×	×	×
		(Kalloubi et al., 2016)	✓	✓	×	×	✓	✓	×	×	×	×
		(Tang et al., 2016)	✓	✓	✓	×	✓	×	×	×	×	×

Category	Section	Reference	NS	Activities		Context		KB	Semantic Web		Entities	
				(R)P	I	P	T		RDF	O	H	URLs
Topic and Event-oriented (cont.)	2.5.2.2	(Subbian et al., 2016)	✓	✓	×	×	×	×	×	×	×	×
		(Zhang et al., 2017)	×	✓	×	✓	×	×	×	×	×	×
		(Yi et al., 2016)	✓	×	×	×	✓	×	×	×	×	×
		(Li et al., 2017)	✓	✓	×	✓	×	×	×	×	×	×
Community Detection	2.5.3	(Barbieri et al., 2013)	✓	✓	×	×	×	×	×	×	×	×
		(Yang et al., 2013)	✓	×	×	×	×	×	×	×	×	×
		(Xia and Bu, 2012)	✓	✓	✓	×	×	×	×	×	×	×
		(Qi et al., 2012)	✓	×	×	×	×	×	×	×	×	×
		(AlFalahi et al., 2013)	✓	×	×	×	×	×	×	×	×	×
		(Slabbekoorn et al., 2016)	✓	✓	×	×	✓	✓	×	✓	×	×
		(Jaitly et al., 2016)	✓	×	×	×	×	×	×	×	×	×
		(Fang and Hu, 2016)	✓	×	✓	×	×	×	×	×	×	×

Appendix B: DBpedia Categories

Table 4.2: Effectiveness of URI Discovery of Phases per Category

Phase	Category	Counter	True Positives	False Positives	%	
					Preci-sion	False Positives
Search by Name	Person	152	94	58	61.84	38.16
	Company	39	24	15	61.54	38.46
	Press	8	6	2	75	25
	Overall	**242**	**156**	**86**	**64.46**	**35.54**
URI Construc-tion	Person	129	109	20	84.5	15.5
	Company	34	32	2	94.12	5.88
	Press	5	5	0	100	0
	Overall	**187**	**164**	**23**	**87.7**	**12.3**

Table 5.9: Origin of Thematic Categories

Thematic Categories				Origin
Political Party	Center	Technology	Basketball	**DBpedia**
Politics	Center Right	Service	Football	
Far Right	Center Left	Sport	Tennis	
Broadcaster	Left Wing	League	Racing	
Artist	Journalist	Soccer	News	
Manufacturer	Telecommunication	Retail	Press	
Economics	Religion	Travel	Leasing	
Athlete	Place	Food	Clothing	
Entertainment	Music	TV	Charity	
Agency	Company	Beverage	Airline	
Right Wing	Openness	Celebrity	Space	**IT**

Table 4.3: Classes and Occurrences per Thematic Category

Thematic Category	Classes	Occurrences	Occurrences per Class	Thematic Category	Classes	Occurrences	Occurrences per Class
Entertainment	4	47	11.75	Food	13	23	1.77
Athlete	3	30	10	Beverage	8	14	1.75
Travel	1	7	7	Service	11	19	1.73
Politics	31	119	3.84	Airline	14	24	1.71
Artist	35	131	3.74	Racing	6	10	1.67
Press	14	49	3.5	Soccer	64	107	1.67
Political Party	12	38	3.17	League	15	23	1.53
Place	7	22	3.14	Agency	6	9	1.5
Music	67	207	3.09	Clothing	4	5	1.25
Company	181	418	2.3	Basketball	19	23	1.21
News	20	46	2.3	Technology	14	16	1.14
Sport	20	45	2.25	Retail	9	10	1.11
TV	33	73	2.21	Charity	3	3	1
Broadcaster	19	38	2	Football	1	1	1
Economics	9	18	2	Leasing	1	1	1
Journalist	6	12	2	Left	1	1	1
Religion	7	14	2	Telecommunication	2	2	1
Manufacturer	34	67	1.97	Tennis	8	8	1
Classes per Thematic Category		**19.5**		**Occurrences per Thematic Category**		**46.67**	**Occurrences per Class**		**2.56**

Appendix C: SPARQL Queries

Query 1: A federated query combining data from the InfluenceTracker and DBpedia services

```
PREFIX it: <http://www.influencetracker.com/ontology#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>

SELECT ?displayName ?influenceMetric ?hIndexRt ?dbpediaUri
?birthDate ?birthPlace ?shortDescription

WHERE {

GRAPH <http://influenceTracker/twitterGraph> {
      <http://www.influencetracker.com/resource/User/barackobama>
            it:hasQualityMetrics ?qm;
            it:hasGeneralInfo ?gi ;
            foaf:account ?account .
      ?qm it:influenceMetric ?influenceMetric ;
            it:hIndexRt ?hIndexRt .
      ?gi it:displayName ?displayName .
}

GRAPH <http://influenceTracker/dbpediaGraph> {
      ?account it:dbpediaUri ?dbpediaUri .
}

SERVICE <http://dbpedia.org/sparql> {
      ?dbpediaUri dbo:birthDate ?birthDate ;
            dbp:birthPlace ?birthPlace ;
            dbp:shortDescription ?shortDescription .
}

}
```

Query 2: Returns the common hashtag counter between two specific users owing Twitter accounts. In our case the first of them is the examined user whose similar accounts will be discovered, while the second one is an arbitrary user to be compared.

```
PREFIX it: <http://www.influencetracker.com/ontology#>

SELECT (COUNT(?ht) AS ?commonHtCounter)

FROM <http://influenceTracker/twitterGraph/full>

WHERE {

<http://www.influencetracker.com/resource/User/{examinedUsername}>
      it:includedHashtag ?ht .
<http://www.influencetracker.com/resource/User/{randomUsername}>
      it:includedHashtag ?ht .

}
```

Query 3: Returns the common domain counter between two specific users owing Twitter accounts. In our case the first of them is the examined user whose similar accounts will be discovered, while the second one is an arbitrary user to be compared.

```
PREFIX it: <http://www.influencetracker.com/ontology#>

SELECT (COUNT(DISTINCT ?domain) AS ?commonDomainsCounter)

FROM <http://influenceTracker/twitterGraph/full>

WHERE {

<http://www.influencetracker.com/resource/User/{examinedUsername}>
      it:includedUrl ?urlExamined .
      ?urlExamined it:domain ?domain .
<http://www.influencetracker.com/resource/User/{randomUsername}>
      it:includedUrl ?urlUser .
      ?urlUser it:domain ?domain .

}
```

Appendix D: Viral Twitter Entities

Table D1: Viral Twitter Entities with top-30% survival rates for January 13, 2014 - seed: “Egypt”

1/8/14	1/9/14	1/10/14	1/11/14	1/12/14	1/13/14	1/14/14	1/15/14
@diet_hawaa	@Morsi_RT	@EgyAntiCoup	@asmaamo83850399	@dignity4theold	@Adamitv	@borzou	@devotion4countr
@el_balad	@youm7	@egypt_now111	@7asnaad1g	@el_balad	@ArabNewsRt7	@dodomoeen	@egypt_now111
@KGirls66	#_	@freedom2mankind	@baladtv	@KGirls66	@arabresistance	@egypt_now111	@el_balad
#_	#abudhabi	@NewIranFree	@CBC_EGY	@Morsi_RT	@Beltrew	@elisferre	@fumi210500576
#abudhabi	#anticoup	@NseejNews	@el_balad	@twahodvoice	@CBC_EGY	@esfahanhanim	@fumikop500579
#android	#bahrain	@rightnowio_feed	@NadaaHesh	@TweetEgyptian	@CBCExtra	@Hazem_Azim	@RiicardoCB
#anticoup	#cairo	@syria_now111	@shamsalhurryah	@youm7	@EgyBloodBank	@IslamRahman	@Shabace1
#ara	#dubai	#abudhabi	@twahodvoice	#_	@egypt_now111	@KarlreMarks	@shadihamid
#arab	#egyconstitution	#anticoup	#_	#bahrain	@el_balad	@kelo3adi	@SonsOfEgypt
#bahrain	#ff	#egyconstitution	#abudhabi	#cairo	@Morsi_RT	@Misr25TV	@syria_now111
#balad	#freethe7	#ff	#ara	#egypt	@TweetEgyptian	@Morsi_RT	@TheBigA7a
#cairo	#iran	#freethe7	#bahrain	#ff	@youm7	@TheMiinz	@youm7
#dubai	#iraq	#iran	#egypt	#freethe7	#_	#_	#_
#egyconstitution	#ksa	#iraq	#ff	#iran	#anticoup	#anticoup	#abudhabi
#ff	#kuwait	#ksa	#freethe7	#iraq	#ara	#bbc	#anticoup
#freethe7	#lebanon	#kuwait	#iran	#ksa	#bahrain	#cairo	#ara
#iran	#morsi	#morsi	#iraq	#kuwait	#balad	#coup	#bahrain
#iraq	#qatar	#saudi	#ksa	#lebanon	#dubai	#coupstitution	#balad
#ksa	#saudi	#sex	#kuwait	#morsi	#egypt	#dubai	#cairo
#kuwait	#sex	#sta	#lebanon	#q8	#ff	#egyconstitution	#dubai
#lebanon	#sisi	#syria	#morsi	#qatar	#freethe7	#egypt	#egypt
#morsi	#syria	#uae	#nowplaying	#rt	#iran	#ff	#ff
#nowplaying	#uae	http://t.co/CQhAHCy4ch	#q8	#saudi	#iraq	#freethe7	#freethe7
#oman	http://t.co/RDEhHMn6tm	http://t.co/D00QkdoG9z	#qatar	#sex	#ksa	#iran	#iran
#q8	http://t.co/sckuvUKUtJ	http://t.co/DWeu2E7E00	#rt	#sta	#kuwait	#iraq	#iraq
#qatar	-	http://t.co/N1GFYmdquR	#saudi	#staracademy	#morsi	#jan25	#ksa
#saudi	-	http://t.co/sckuvUKUtJ	#sta	#syria	#news	#ksa	#kuwait
#sex	-	http://t.co/WG2aABrX8g	#staracademy	#uae	#rt	#kuwait	#lebanon
#syria	-	-	#syria	http://t.co/3eoTyXd1rB	#saudi	#morsi	#qatar
#uae	-	-	#uae	http://t.co/D00QkdoG9z	#sharon	#news	#referendum
http://t.co/DWeu2E7E00	-	-	http://t.co/3eoTyXd1rB	http://t.co/DWeu2E7E00	#sta	#obama	#saudi
http://t.co/J9xh2TYLNt	-	-	http://t.co/DWeu2E7E00	http://t.co/J9xh2TYLNt	#syria	#r4bia	#sta
http://t.co/lKh2rCWL0V	-	-	http://t.co/H0EqkoVsG2	http://t.co/KswIwWWaoI	#uae	#referendum	#syria
-	-	-	http://t.co/J9xh2TYLNt	-	#un	#saudi	#uae
-	-	-	http://t.co/RDEhHMn6tm	-	#usa	#syria	#un
-	-	-	-	-	http://t.co/sckuvUKUtJ	#uae	#yemen
-	-	-	-	-	http://t.co/yb4c67OkOJ	http://t.co/sckuvUKUtJ	http://t.co/lKh2rCWL0V
-	-	-	-	-		http://t.co/yb4c67OkOJ	http://t.co/ORk0I6FSl2
-	-	-	-	-	-	http://t.co/ZdhnOToith	http://t.co/RDEhHMn6tm
-	-	-	-	-	-	-	-

Table D2: Viral Twitter Entities with top-20% survival rates for January 13, 2014 - seed: “Syria”

1/8/14	1/9/14	1/10/14	1/11/14	1/12/14	1/13/14	1/14/14	1/15/14
@Free_Media_Hub	@GSolaimani	@AM000Z	@Free_Media_Hub	@30mehr	@Free_Media_Hub	@Free_Media_Hub	@melissarfleming
@mxxxx_mmmmm7430	@tintin1957	@epaulnet	@GeredaV	@Free_Media_Hub	#abc	@IAC4RC	@nasrinforiran
@nasrinforiran	#_	@greyfoxguy	@GSolaimani	@GSolaimani	#afp	@nasrinforiran	@SyriaTwitte
@Shareif	#aleppo	@mxxxx_mxmxm7422	@i_magpie	@SIranChange	#aljazeera	@Refugees	@UN
@tintin1957	#aljazeera	@nasrinforiran	@NcrIran	@tintin1957	#ap	#_	#_
#_	#assad	@no2censorship	@RevolutionSyria	#_	#assad	#abc	#afp
#aleppo	#bahrain	@TheIslamicUmmah	@tintin1957	#abc	#campashraf	#afp	#aleppo
#aljazeera	#campashraf	@tintin1957	#abc	#afp	#campliberty	#aljazeera	#aljazeera
#assad	#cbs	#abc	#aljazeera	#aljazeera	#cnn	#assad	#assad
#bahrain	#foxnews	#ap	#ap	#bahrain	#egypt	#breaking	#cnn
#breaking	#freethe7	#assad	#assad	#breaking	#freethe7	#campashraf	#egypt
#campashraf	#google	#bahrain	#assadcrimes	#damascus	#gasattacks	#campliberty	#euronews
#cbs	#hama	#breaking	#bahrain	#fox	#geneva2	#cnn	#foxnews
#damascus	#homs	#campashraf	#belgium	#freethe7	#google	#damascus	#freethe7
#egypt	#iran	#cnn	#campashraf	#health	#iran	#egypt	#google
#freethe7	#iraq	#egypt	#campliberty	#iran	#iraq	#euronews	#health
#fsa	#isis	#fox	#damascus	#iraq	#isis	#fox	#iran
#gasattacks	#kuwait	#freethe7	#euronews	#isis	#jordan	#freethe7	#iraq
#google	#lebanon	#gasattacks	#fox	#islam	#lebanon	#google	#kuwait
#iran	#nbc	#homs	#freethe7	#nbc	#nbc	#iran	#nbc
#iraq	#news	#iran	#google	#news	#pmoi	#iraq	#pmoi
#ksa	#politics	#isis	#homs	#sms	#politics	#lemonde	#rajavi
#lebanon	#qatar	#london	#iran	#sun	#reuters	#libyan	#reuters
#nbc	#rajavi	#london	#ksa	#sydney	#russia	#nbc	#saudi
#news	#reuters	#news	#kuwait	#syria	#seckerry	#pmoi	#sms
#oman	#russia	#pmoi	#lebanon	#un	#syria	#rajavi	#sun
#politics	#sms	#rajavi	#lemonde	#unami	#uae	#reuters	#sydney
#sms	#syria	#reuters	#unami	#usa	#un	#syria	#syria
#sun	#uae	#rt	#unhcr	http://t.co/5Jsm2IRg0C	#unami	#un	#syriacrisis
#syria	#uk	#sms	#upi	http://t.co/vUZtxXoomC	#upi	#unhcr	#uae
#un	#un	#sun	#usa	-	#world	#world	#un
#unami	#unami	#syria	#world	-	http://t.co/01tSxI2klt	http://t.co/1J2mrY70Fq	http://t.co/SoBQgWybzJ
#unhcr	#upi	#us	http://t.co/8KabhVygS9	-	http://t.co/6JEZvw5HyR	http://t.co/d31Qw0p0uH	http://t.co/YZvwPljb7y
#usa	#world	#usa	http://t.co/krvCS7rs2d	-	http://t.co/83SrzqVlpn	http://t.co/r21jSS4Eah	-
#world	http://t.co/i8QegFximG	http://t.co/1E9UxXC3I9	-	-	http://t.co/jhGROR01IV	-	-
http://t.co/8nhfTJO9iz	http://t.co/KssepuNIfV	http://t.co/8N1W4UhsD0	-	-	-	-	-
http://t.co/F79vjGXToI	-	http://t.co/bTYwjRs393	-	-	-	-	-
	-		-	-	-	-	

Appendix E: Publications

Table E1: Journal articles

(Anagnostopoulos et al., 2015)	Anagnostopoulos, I, Razis, G, Mylonas, P and Anagnostopoulos, CN 2015, 'Semantic query suggestion using Twitter Entities', *Neurocomputing*, vol. 163, pp. 137-150. https://doi.org/10.1016/j.neucom.2014.12.090
(Razis and Anagnostopoulos, 2016)	Razis, G and Anagnostopoulos, I 2016, 'Discovering similar Twitter accounts using semantics', *Engineering Applications of Artificial Intelligence*, vol. 51, pp. 37-49. http://dx.doi.org/10.1016/j.engappai.2016.01.015
(Razis et al., 2018)	Razis, G, Anagnostopoulos and I, Sherali, Z 2018, 'Modeling Influence With Semantics in Social Networks: a Survey', *ACM Computing Surveys (under review)*, draft version at https://arxiv.org/abs/1801.09961

Table E1: Conference Proceedings

(Anagnostopoulos et al., 2013)	Anagnostopoulos, I, Razis, G, Mylonas, P and Anagnostopoulos, CN 2013, 'Query Expansion with a Little Help from Twitter', *Engineering Applications of Neural Networks, Communications in Computer and Information Science*, vol. 384, pp. 20-29. http://dx.doi.org/10.1007/978-3-642-41016-1_3
(Razis and Anagnostopoulos, 2014a)	Razis, G and Anagnostopoulos, I 2014, 'InfluenceTracker: Rating the impact of a Twitter account', *Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology*, vol. 437, pp. 184-195. https://doi.org/10.1007/978-3-662-44722-2_20
(Razis and Anagnostopoulos, 2014b)	Razis, G and Anagnostopoulos, I 2014, 'Semantifying Twitter: the InfluenceTracker Ontology', *in 9th International Workshop on Semantic and Social Media Adaptation and Personalization*, Corfu, pp. 98-103, http://dx.doi.org/10.1109/SMAP.2014.23
(Razis et al., 2015)	Razis, G, Anagnostopoulos, I and Vafopoulos, M 2015, 'Semantic social analytics and Linked Open Data cloud', *in 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Trento, pp. 1-6. http://dx.doi.org/10.1109/SMAP.2015.7370091
(Razis et al., 2016)	Razis, G, Anagnostopoulos and I, Saloun, P 2016, 'Thematic labeling of Twitter accounts using DBpedia properties', *in 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Thessaloniki, pp. 106-111. http://dx.doi.org/10.1109/ SMAP.2016.7753393