Foundations and Trends® in Information Retrieval

Extracting, Mining and Predicting Users' Interests from Social Media

Suggested Citation: Fattane Zarrinkalam, Stefano Faralli, Guangyuan Piao and Ebrahim Bagheri (2020), "Extracting, Mining and Predicting Users' Interests from Social Media", Foundations and Trends® in Information Retrieval: Vol. 14, No. 5, pp 445–617. DOI: 10.1561/1500000078.

Fattane Zarrinkalam
Ryerson University
Canada
fzarrinkalam@ryerson.ca

Stefano Faralli
University of Rome Unitelma Sapienza
Italy
stefano.faralli@unitelmasapienza.it

Guangyuan Piao
Insight Centre for Data Analytics, NUI Galway
Ireland
guangyuan.piao@insight-centre.org

Ebrahim Bagheri
Ryerson University
Canada
bagheri@ryerson.ca

now
the essence of knowledge
Boston — Delft

Contents

Extracting, Mining and Predicting Users' Interests from Social Media

Fattane Zarrinkalam[1], Stefano Faralli[2], Guangyuan Piao[3] and Ebrahim Bagheri[4]

[1] *Ryerson University, Canada; fzarrinkalam@ryesron.ca*
[2] *University of Rome Unitelma Sapienza, Italy; stefano.faralli@unitelmasapienza.it*
[3] *Insight Centre for Data Analytics, NUI Galway, Ireland; guangyuan.piao@insight-centre.org*
[4] *Ryerson University, Canada; bagheri@ryesron.ca*

ABSTRACT

The abundance of user generated content on social media provides the opportunity to build models that are able to accurately and effectively extract, mine and predict users' interests with the hopes of enabling more effective user engagement, better quality delivery of appropriate services and higher user satisfaction. While traditional methods for building user profiles relied on AI-based preference elicitation techniques that could have been considered to be intrusive and undesirable by the users, more recent advances are focused on a non-intrusive yet accurate way of determining users' interests and preferences. In this monograph, we will cover five important subjects related to the mining of user interests from social media: (1) the foundations of social user interest modeling, such as information sources, various types of representation models and temporal features, (2) techniques that have been adopted or proposed for

Fattane Zarrinkalam, Stefano Faralli, Guangyuan Piao and Ebrahim Bagheri (2020), "Extracting, Mining and Predicting Users' Interests from Social Media", Foundations and Trends® in Information Retrieval: Vol. 14, No. 5, pp 445–617. DOI: 10.1561/1500000078.

mining user interests, (3) different evaluation methodologies and benchmark datasets, (4) different applications that have been taking advantage of user interest mining from social media platforms, and (5) existing challenges, open research questions and opportunities for further work.

1

Introduction

Mining user interests from user behavioral data is critical for many applications, ranging from homophily analysis to recommender systems. Based on user interests, service providers such as advertisers can significantly reduce service delivery costs by offering the most relevant products (e.g., ads) to their customers. The challenge of accurately and efficiently identifying user interests has been the subject of increasing attention in the past several years (Zarrinkalam *et al.*, 2019a). Early approaches were based on explicit input from individuals about their own interests (Maron *et al.*, 1986). To avoid the extra burden of manually filling in and maintaining interest profiles, most methods in the past two decades have focused on the development of techniques that can automatically and unobtrusively determine users' interests based on user behavioral data from data sources such as browsing history, page visits, the links they click on, the searches they perform and the topics they interact with (Gasparetti, 2017; Holub and Bieliková, 2010; Li and Zhang, 2013).

With the emergence and growing popularity of social media such as blogging systems, wikis, social bookmarking, social networks and microblogging services, many users are extensively engaged in at least some of these applications to express their feelings and views about a wide variety of social events/topics as they happen in real time by

commenting, tagging, joining, sharing, liking, and publishing posts (Abel *et al.*, 2011b; Li *et al.*, 2008). According to Statista, a company which provides statistics and survey results, there were an estimated 3.6 billion people using social media in 2020, a number projected to increase to almost 4.41 billion in 2025.[1] This has made social media an exciting and unique source of information about users' interests.

For instance, when looking at Twitter data during the first week of March 2019, the rivalry between the two English Premier League soccer clubs, i.e., *Tottenham Hotspur* and *Arsenal*, is a topic that has attracted a lot of discussion and interest. The development of techniques that can automatically detect such topics and model users' interests towards them has the potential to improve the quality of applications that work on a user modeling basis, such as filtering twitter streams (Kapanipathi *et al.*, 2011), news recommendation (Abel *et al.*, 2011b) and retweet prediction (Feng and Wang, 2013), among others.

This monograph is a valuable resource for those who have familiarity with social media mining and basics of information retrieval (IR) techniques. Where appropriate, the monograph will not make any assumptions about the researchers' knowledge on more advanced techniques such as link prediction, matrix factorization, entity linking and knowledge graph-based reasoning, among others. As such, sufficient details about user interest modeling from social media will be provided as appropriate so that the content will be accessible and understandable to those who have fundamental understanding of IR principles. The monograph will only assume familiarity with topics included in an undergraduate IR course such as those covered in Manning *et al.* (2008).

The monograph is structured as follows.

- Section 1 provides the motivations of user interest modeling from social media and the scope and delimitation of this monograph by highlighting the difference between this monograph and other related review monographs and introducing some related research areas which are out of the scope of this monograph.

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

- Section 2 introduces the foundations of user interest modeling from social media such as information sources, representation units to represent each topic of interest and user interest profile, temporal aspects and cross-system user interest modeling.
- Section 3 describes user interest modeling approaches by focusing on three main perspectives: (1) explicit user interest detection, (2) implicit user interest mining, and (3) future user interest prediction.
- Section 4 describes the main evaluation methodologies which have been widely adopted in the literature followed by the existing benchmark datasets and evaluation metrics.
- Section 5 introduces different applications that have been taking advantage of user interest modeling from social media platforms to improve their services.
- Section 6 presents exciting open challenges, future directions and research questions in the state-of-the-art for modeling users' interests from social media.

1.1 Definitions

In this section, we provide concrete definitions of *social media* and *user interest modeling* as two key terms used in this monograph.

Social Media. The term *social media* has been defined in the literature in various ways by different communities such as communications, management, and computer science (Boyd and Ellison, 2007; Fuchs, 2014; Rohani and Hock, 2010). In this monograph, we follow the same definition provided by Obar and Wildman (2015), which synthesize the definitions presented in the literature by defining *social media* as a service that has the following four distinct commonalities:

- *Social media services are (currently) Web 2.0 Internet-based applications.* In Web 2.0 applications, users have become not only content consumers but also active producers.

- *User-generated content is the lifeblood of social media.* Social media services are not sustainable without user-generated content. For example, the videos that we upload to YouTube, the connections and posts that we generate on Facebook or Twitter play crucial roles in making those social media platforms live.

- *Individuals and groups create user-specific profiles for a site or application designed and maintained by a social media service.* User profile in a social media service provides a unique way of identifying each user which is important to enable social networks between users. A user profile here can refer to any information to uniquely identify a user such as a username, IP address, locations, contact information, etc.

- *Social media services facilitate the development of social networks online by connecting a profile with those of other individuals and/or groups.* For example, friends on Facebook, connections on LinkedIn and followers on Twitter and Instagram. Users are motivated to create their social networks in social media services for consuming content generated by their social networks or interact with them.

User Interest Modeling. Piao and Breslin (2018a) provided a general definition about user interest modeling and user profiles. We use a refined definition as follows.

The process of obtaining the user interest profile is called user interest modeling. A user interest profile is a data structure that represents the degree of interest of an individual user over a set of topics represented by words or concepts.

1.2 Related Review Papers

Despite the importance of user interest modeling from social media, there is a lack of an extensive review on this domain that covers the ideas, insights and applications of different approaches in user interest modeling from social media. Piao and Breslin (2018a) have reviewed studies in user interest modeling from microblogging websites such as Twitter by focusing on four dimensions: (1) data collection, (2) representation of

user interests, (3) construction and enhancement of user interest profiles, and (4) evaluation methodologies.

The authors in Piao and Breslin (2018a) have presented the foundations of user interest modeling from microblogging websites and overlooked other social media platforms (e.g., Facebook, Flickr and Pinterest). Further, they have not covered the potential applications of the extracted user interest models. Since we categorize different approaches proposed in the literature with respect to three main perspectives: (1) explicit user interest detection, (2) implicit user interest mining, and (3) future user interest prediction, this has set the stage for a more detailed presentation of the ideas and insights about different user interest modeling approaches. Moreover, in addition to providing the fundamental information that new researchers need for understanding this field, we extensively discuss the potential applications of the extracted user interest models and promising techniques which can be applied for future exploration in this field.

As another related review monograph, Safari *et al.* (2019) have recently reviewed different studies on user behavior mining from social media (UBMSM). They have provided different statistical and demographic information (e.g., venue types and publishers) about the publications in this domain. Further, based on the focus area of their studies, they have identified four main aspects which affect user behavior mining from social media: (1) user, (2) content, (3) network structure, and (4) information diffusion. For each aspect, multiple characteristics are defined and their impact and consequences on UBMSM are discussed. The authors have considered user interest modeling as one of the four characteristics of the user aspect and reviewed different related studies on this topic.

Therefore, in contrast to our work, the focus of the authors in Safari *et al.* (2019) is not directly on user interest modeling from social media and they have concentrated more on the effect of different characteristics and aspects of user behavior. As a result, they don't provide more in-depth analysis and discussion of the studies in this field. However, our monograph describes the specific techniques, evaluation strategies, benchmarks and challenges of user interest modeling from social media, in addition to concrete directions for future work. Consequently, we

believe that our review monograph is more insightful for a researcher interested in this specific field, i.e., user interest modeling from social media.

1.3 Related Research Areas

There are some research areas/topics which are related to user interest modeling from social media. Because each of these areas is a mature and active field of research and includes a rich line of studies in the literature, we will not cover them in detail and they are beyond the scope of this monograph. To provide a more in-depth analysis, we only review the studies that focus directly on user interest modeling from social media. In the following, some of the most important examples of these areas/topics, e.g., topic detection, personality prediction and latent user modeling from social media are introduced and their similarities and differences to the subject of this monograph are highlighted.

1.3.1 Topic/Event Detection from Social Media

There is already a well-established body of work in the literature that extracts topics/events from social media (Aiello *et al.*, 2013; Huang *et al.*, 2017a; Petkos *et al.*, 2014; Yan *et al.*, 2015). Applying topic modeling methods, such as LDA, over social posts is the main approach to extract topics from social media. However, since the majority of standard topic modeling methods are designed for regular documents such as news articles, they fail to identify the essential information of social posts which are short, noisy and informal. An intuitive solution to address this issue is first using a *pooling scheme* to aggregate the related social posts to a single document (e.g., posts published by a given user or in a given time interval) and then applying a standard topic modeling method on the resulting documents to extract topics from social posts. This allows for the discovery of better topics without modifying the existing topic modeling methods (Alvarez-Melis and Saveski, 2016; Mehrotra *et al.*, 2013; Rajani *et al.*, 2014).

To extract topics from social posts, instead of applying pooling scheme, some studies have applied some restrictions to simplify the

conventional topic models or developed novel topic models. For example, in Zhao *et al.* (2011), Twitter-LDA model has been proposed which assumes that a single tweet contains only one topic. Similarly, in Yan *et al.* (2015), the authors have extended the Biterm Topic Model (BTM) (Yan *et al.*, 2013), which models word pairs instead of words for effective topic modeling in short texts, by incorporating the burstiness of word pairs as prior knowledge in BTM for bursty topic modeling from social posts.

There is also another line of studies that extracts topics/events from social media by applying clustering methods over social posts or their features (Comito *et al.*, 2019a,b; Long *et al.*, 2011). As one of the earlier studies that focused on microblogging data, in Long *et al.* (2011), a *co-occurrence graph* is constructed by extracting topical words from daily posts. To extract events during a time period, the authors have applied a top-down hierarchical clustering algorithm over the co-occurrence graph.

The above studies can be considered as a related area to user interest modeling as they are applied in some studies to first extract active topics in social media and then the interest profile of users are modeled over the extracted topics (Arabzadeh *et al.*, 2018; Zarrinkalam *et al.*, 2018). In these studies, it is assumed that existing state-of-the-art techniques can be employed for extracting and modeling topics. Therefore, they are not engaged with proposing a new method for the identification of topics and only have focused on determining the degree of interest of users towards the topics once they are identified. Given this focus, we review the work related to the problem of user interest modeling from social media. Interested readers are encouraged to see Bhardwaj *et al.* (2019), Farzindar and Khreich (2015), Zarrinkalam and Bagheri (2016) for the state of the art on topic and event detection from social media.

1.3.2 Personality Prediction of Social Media Users

Predicting the personality of users from social media is another related research area to user interest modeling. In Kosinski *et al.* (2015), the authors have provided a comprehensive discussion about opportunities,

challenges and ethical aspects of social media involvement in psychological studies. As an example, they have studied users' profiles in social media in order to perform personality classification. Similarly, in Youyou *et al.* (2015), the authors have concluded that digital footprints created on top of user preferences (i.e., Facebook likes) are more accurate and valid than judgments made by social-cognitive experts. Therefore, there is already a well-established body of work in the literature that predicts users' personality traits from social media (Golbeck *et al.*, 2011; Souri *et al.*, 2018).

Most studies in this area have used the *Big Five* model (Halverson, 1994) as one of the most well-known measures for personality traits (the five personality traits according to *Big Five* model are openness, conscientiousness, extroversion, agreeableness and neuroticism). For example, in Souri *et al.* (2018), the authors have adopted the *Big Five* model to design a classifier which is able to automatically identify five classes of personality traits based on the users' social media profiles. In Golbeck *et al.* (2011), social behavior of a user is considered as a strong indicator to predict her personality. To model the behavior of a user in social media, the authors have developed a set of behavioral features based on the intensity and number of social interactions that the user has with her friends along a number of dimensions such as reciprocity and priority. They have analyzed these behavioral features along with a set of textual features and showed that behavioral features perform as well as textual features for predicting user personality.

Similar to user interest modeling approaches, the above studies try to model the users' behavior on social media by mainly analysing the textual content of users and their social relations. However, since these studies aim at extracting personality traits of users instead of extracting the users' interests, they dig more into the linguistic features. For example, to identify the personality of users, in many studies LIWC (Pennebaker *et al.*, 1999), which produces statistics on 81 different linguistic features of text, is utilized to study various emotional, cognitive, structural, and process components presented in the users' textual content. Interested readers about the state-of-the-art approaches on personality identification from social media are encouraged to see Kaushal and Patwardhan (2018).

1.3.3 Latent User Modeling from Social Media

There are plenty of studies that embed a user's information in social media such as the user's relations, textual content and demographic profile, into a latent low dimensional space (Benton *et al.*, 2016; Zhang *et al.*, 2017). These studies are usually tied to a specific task from which the model is learned and they represent user models by latent features which are not human readable.

For example, in Ding *et al.* (2017), the user's information on Facebook (e.g., *likes* and *status updates*) is embedded in order to build substance use detection systems to identify users who are at risk of substance use disorder. The authors have employed Doc2Vec (Le and Mikolov, 2014), which produces a dense low dimensional feature vector for a document, as one of their approaches to embed users. To learn user embeddings from social media posts, they introduced two methods: (1) User-D2V which treats all the posts published by each user as one document and trains a document vector to represent each user. (2) Post-D2V which learns a vector for each post and then aggregates all the post vectors from the same user to derive the user embedding.

As other examples, in Benton *et al.* (2016), the authors have proposed to embed different information of a user on Twitter (e.g., the user's posts, followers and friends) into a single embedding vector by applying a multiview approach. They have shown the effectiveness of their model on three different prediction tasks (i.e., user engagement prediction, friend recommendation and demographic characteristics inference). Word2Vec (Mikolov *et al.*, 2013) is one of the methods the authors have applied to represent each view of the user by simply averaging the word embeddings for all the words within that view (e.g., the user's posts). In Piao and Breslin (2018b), the latent representation of social posts and users on Twitter are learned for the application of tweet recommendation. Their proposed model employs Long Short-Term Memory neural networks (LSTMs) (Piao and Breslin, 2018b) for learning tweet embeddings, and calculates the degree of interest of a user to a tweet based on the similarity between the user and the tweet embeddings as well as the similarity between the user and tweet's publisher embeddings.

In this monograph, our focus is on reviewing the user interest modeling approaches that identify the degree of interest of a user over a set of topics each of which are represented by words or concepts. In other words, we do not review user modeling approaches that result in latent user models represented by a vector of numbers which are not human readable. A recent survey about social media-based user embedding can be found in Pan and Ding (2019).

2

Foundations of Social User Interest Modeling

In this section, we introduce the fundamental information that new researchers need for understanding user interest modeling from social media such as different information sources, user interest representation units, temporal user interest modeling, semantics-enabled user interest profile representation models, and cross-system user interest modeling. In order to help the readers better understand the terminology used in the context of user interest modeling from social media, we have provided a glossary of terms in the Appendix.

2.1 Information Sources

In this section, we discuss various information sources in online social media for collecting data in order to infer user interest profiles. They can be divided into two main categories: (1) *internal data* which are collected from the existing information on online social media such as textual content of users or their social connections, and (2) *external data* which are collected from other sources of information such as knowledge bases in order to improve the process of user interest modeling.

It is noted that, due to the public nature of Twitter content and its privacy policies[1] which enables researchers to easily collect and process data, Twitter is the most popular social media platform in the literature. However, although most of the studies on user interest modeling mainly focus on different information sources on Twitter, in the following sections, we also cover other social media such as Instagram, Facebook and Tumblr.

2.1.1 Internal Data

Users on social media can explicitly describe themselves in their user profiles. Therefore, information provided by a user by filling her profile fields is an explicit source of information for inferring the user's interests. Biography (or simply *bio*) is a user profile field in most social media (e.g., Twitter, Instagram and Facebook) which includes informative information for user interest modeling (Piao and Breslin, 2017a). For example, based on the biography of a user on Twitter – *Fan of Arsenal and Real Madrid*, one can infer that the user is interested in Arsenal or Real Madrid. As other examples, *job titles*, *education*, *skills* and *interests* are some fields in the LinkedIn profile about a user that can be used to infer her interests (Piao and Breslin, 2016b). However, the information in user profiles are not always completely or accurately provided by the users. Further, since these fields are usually answered by the users during the registration process and it is unlikely that the users edit this information afterwards, inferring user interests from user profiles cannot automatically adapt to shifts in users' interests.

Therefore, the main source of information for inferring users' interests from social media is the *social posts* that users interact with (e.g., by liking, publishing, sharing). A user's social posts usually include a set of terms that shows the user's topics of interest. For example, one can infer that a user is interested in Arsenal if the user has mentioned the term "Arsenal" frequently in her posts. Instead of using low-level features such as term tokens, in some studies higher-level features such as named entities or semantic concepts are extracted from the textual

[1] https://developer.twitter.com/en/developer-terms/agreement-and-policy.html.

content of posts and considered as the user's topics of interest (Zhao *et al.*, 2015).

Some social posts have embedded links (*URL*s) or *tags* to include more information about their topic. For example, it is shown that 66.6% of Instagram posts contain a hashtag.[2] Therefore, *URL*s and *tags* can also be considered as two informative information sources for identifying interests and intention of a user (Peña *et al.*, 2013; Piao and Breslin, 2016d). Further, in the last several years, with the emergence of different types of social media platforms such as photo-centric ones (e.g., Flickr, Instagram and Tumblr), multimedia content such as photos and videos attached to social posts are utilized as another important information source for mining user interests from social media (Joshi *et al.*, 2015). For example, Grbovic *et al.* (2016b) showed that 78.11% of posts in Tumblr contain images.

Instead of or in addition to analyzing social posts to model user interests, some studies have proposed using the *social relations* of users (Myers and Leskovec, 2014). For example, on Twitter, a user is allowed to explicitly *follow* other users without their permission (the user who follows a person is called a "follower", while the one who is followed is called "friend"). Therefore, by analysing social relations, one can infer that a user is interested in Arsenal if she is following the Twitter account *@Arsenal*. Further, based on the idea that a user's behaviors are affected by opinions of their friends, a user's interests can be identified by considering the interests of her friends (Bao *et al.*, 2013). For example, on Sina Weibo,[3] the most popular microblogging platform in China, if a user has many friends who are interested in Arsenal, it is probable that the user is also interested in Arsenal.

In addition to explicit relations between users in social media (e.g., friendship or followership), there are some implicit relations. A user can *share* a post published by a given user or *reply/comment* to it if she is interested in the content of that post. Further, a user is allowed to *mention* another user in her posts. Therefore, in some studies, these

[2] https://mention.com/en/blog/instagram-statistics-report/.
[3] https://www.weibo.com/login.php/.

implicit relations between users are also considered as indicators for inferring user interests (Wang *et al.*, 2013).

As another valuable source of information for user interest modeling, several studies have recently leveraged *List membership* of followees on Twitter (Piao and Breslin, 2017b). Users on Twitter can freely create a topical list and add other users who are expert in that topic or have been frequently tweeting about news on the topic. For instance, a user might create a list membership named "Football" and add fans of football to the list. Therefore, if a user is following the users who have been added into many topical lists related to the topic of football, one can infer that the user is interested in football.

2.1.2 External Data

Social posts are the main source of information for inferring a user's interests. However, since social posts such as tweets are often short, noisy and informal, inferring user interests from social posts is a challenging task. To address this challenge, some studies have explored how to leverage the information in external knowledge bases to enrich the content of social posts (Bontcheva and Rout, 2014). For example, it is shown that more than 85% of posts on Twitter are related to news (Kwak *et al.*, 2010). Therefore, in some studies, social posts are linked to news articles and their content is enriched with the content of news articles (Abel *et al.*, 2011c). However, since linking posts to news articles requires maintaining up-to-date news streams from mainstream news providers such as CNN,[4] BBC[5] and New York Times,[6] some studies have simply leveraged the content of the embedded URLs in posts to enrich them and provide additional information about the post (Piao and Breslin, 2016d).

Knowledge bases such as Wikipedia, Freebase and WordNet are other external sources of information that are incorporated by some researchers to enhance the accuracy of user profiles. For example, some

[4] https://www.cnn.com.
[5] https://www.bbc.com.
[6] https://www.nytimes.com.

studies have proposed to enrich posts by annotating them with unambiguous semantic concepts described in external knowledge bases. These knowledge bases provide explicit semantic description of concepts and their relationships. Therefore, they can provide additional contextual information about social posts and their underlying semantics (Kapanipathi *et al.*, 2014; Michelson and Macskassy, 2010).

2.2 User Interest Representation Unit

Regardless of which information sources are used for inferring user interests from social media, in this section we introduce how each topic of interest has been represented in different approaches. There are two main user interest representation units: (1) keyword-based, and (2) concept-based models.

2.2.1 Keyword-Based Models

A simple and predominant approach for representing a user's interests which is still popular in information filtering systems is using keywords (unigrams | #tags) mentioned in the textual content of the user to represent her topics of interest. There is a line of work that represents each topic of interest as a *single keyword*. However, since users in social media can freely publish posts without any restriction, their posts are usually unstructured and include a nearly unlimited set of keywords. Therefore, representing each topic of interest by a keyword results in a sparse representation of a user's interests. Further, since keyword-based approach forgoes the underlying semantics of textual content, representing each topic of interest as a single keyword, it suffers from well-known problems in natural language processing such as polysemy and synonymy (Chen *et al.*, 2010; Yang *et al.*, 2012).

For example, given the sample tweets in Table 2.1, Arsenal keyword is mentioned in both tweets but with different meanings. In the first tweet, Arsenal refers to *Arsenal Football Club*; while in the second tweet, it refers to a 2017 *American film* directed by Steven C. Miller and written by Jason Mosberg. By representing each user's topic of interest by a keyword, the users who are engaged with each of these tweets will

Table 2.1: Two sample tweets including Arsenal keyword

	Tweet Content
1	Arsenal won't win with Wenger's policy. Spurs continue to exceed expectations...
2	Here's my review of #Arsenal, otherwise known as the first Nicolas Cage film of 2017 http://substreammagazine.com/2017/01/arsenal-movie-review/

be identified to have the same interests, i.e., Arsenal, while they have completely different interests.

There is another keyword-based representation model, named *group of keywords*, in which each topic of interest has associated to a set of keywords with their probabilities with respect to the topic and usually distilled from topic modeling approaches such as Latent Dirichlet Allocation (LDA) (Blei, 2012). Since the group of keywords representation model considers a combination of related keywords as a topic, to some extent, it overcomes some limitations of single keyword representation models such as polysemy and synonymy.

Two sample LDA topics are illustrated in Figure 2.1. The first topic that refers to Arsenal football club includes *Arsenal*, *League*, *Cups*, *Club* and *Football* as its main keywords and the main keywords in the second topic which refers to Arsenal film are *Arsenal*, *American*, *Film*, *2017* and *Mikey*. Therefore, based on the constituent keywords in each topic, it is possible to infer the meaning of Arsenal. However, since the topics are still built on keywords, they don't consider the underlying semantics of text. Further, since the topic modeling approaches such as LDA designed for regular documents such as news, they may not work well on social posts which are noisy and informal and might suffer from sparsity problem (Sriram *et al.*, 2010).

Two main approaches are introduced in the literature for applying LDA to extract user interests as a distribution over topics: (1) user-LDA, and (2) post-LDA. In user-LDA, the author-topic model (Steyvers *et al.*, 2004), which is a generative model that extends the LDA approach to include authorship information, is applied to social posts. In fact, all the posts of a user are aggregated as a single document and then

Figure 2.1: The extracted LDA topics for *Arsenal Football Club* (left) and *Arsenal Movie* (right).

LDA is applied to infer topics and the degree of users' interests over the extracted topics (Hong and Davison, 2010; Mehrotra *et al.*, 2013). However, in post-LDA, each post is considered as a single document and then LDA is applied. Finally, the interest profile of a user is inferred by aggregating the topic distribution vectors of her published posts. In Ding *et al.* (2017), it is shown that the post-LDA approach learns a better user representation model compared to user-LDA in the underlying applications, because in social media a small number of users usually accounts for a large amount of the content (Antelmi *et al.*, 2019).

To improve the performance of the user-LDA model to extract user interests, there are some studies that modify the author-topic model by introducing some assumptions. For example, in Xu *et al.* (2011) the authors have proposed Twitter-user model as an extension of the author-topic model by assuming that each tweet is associated with a latent variable that indicates whether the tweet is related to its author's interest. Similarly, Twitter-LDA has been proposed in Zhao *et al.* (2011) which assumes that a single tweet contains only one topic. Since Twitter-LDA is not able to do online inference and does not consider dynamics of user interests, in Sasaki *et al.* (2014), the authors have extended Twitter-LDA by (1) estimating the ratio between topic words and general words for each user to model the generation process

of tweets more accurately, and (2) estimating the dynamics of user interests and topic trends online based on the topic tracking model.

2.2.2 Concept-Based Models

To address some problems of keyword-based representation models, in some studies, concepts from knowledge bases such as Wikipedia/DBpedia[7] are utilized to represent user interests. They usually represent each topic of interest by a single concept. Two main types of concepts used in different studies are *entity* and *category*. There are also some studies that have utilized both types for representing users' interests which are called *hybrid* models.

In these studies, to represent each topic of interest as an *entity*, first an entity linking method is applied to extract entities from textual content of social posts and then instead of modeling the users' interests using the keywords, the identified entities are utilized to represent user interests (Orlandi *et al.*, 2012; Zarrinkalam *et al.*, 2018). The primary goal of entity linking of textual content is to process a textual document, identify the mentions that have the potential to be linked to some entities in knowledge bases such as DBpedia or Freebase and connect them to those entities. This allows for text analytics at a higher level which focuses on the meaning of the documents in addition to its syntactics (Feng *et al.*, 2018; Shen *et al.*, 2013).

Early work on entity linking has focused on long and formal texts and considered contextual and semantic similarities between the text and candidate entities in knowledge bases (Ceccarelli *et al.*, 2013; Mihalcea and Csomai, 2007; Milne and Witten, 2008). However, Recent studies in this area have shifted to entity linking of texts with special characteristics such as search queries and social posts which is challenging due to the lack of context (Ferragina and Scaiella, 2012; Hasibi *et al.*, 2016; Meij *et al.*, 2012; Reinanda *et al.*, 2015). It is noted that, the majority of studies in the field of user interest modeling from social media have extracted entities mentioned in the textual content of social posts by

[7] DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia (http://dbpedia.org).

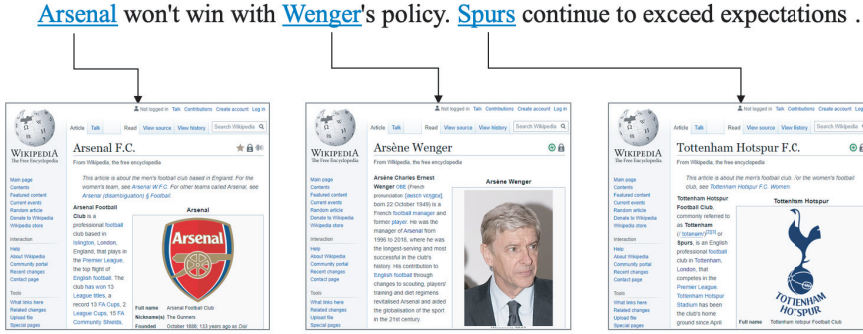

Figure 2.2: Extracted Wikipedia entities for tweet *"Arsenal won't win with Wenger's policy. Spurs continue to exceed expectations…"*.

utilizing the existing entity linking tools (e.g., TagMe,[8] Zemanta[9] and Aylien[10]) instead of proposing new methods.

Similarly, for photo-centric social media platforms, in order to represent each topic of interest as an entity, most studies first use one of well-established computer vision approaches such as ResNet-50 (He *et al.*, 2015a) to classify photos into the 1000 ImageNet concepts (Russakovsky *et al.*, 2015). Then, those ImageNet concepts are mapped to a corresponding entity in a knowledge base such as DBpedia or BabelNet.[11]

Figures 2.2 and 2.3 show two sample tweets and their extracted Wikipedia entities. For example, the first tweet, i.e., "*Arsenal won't win with Wenger's policy. Spurs continue to exceed expectations*…", will be modeled as a collection of three entities, namely Arsenal_F.C,[12] Arsene_Wenger[13] and Tottenham_Hotspur_F.C.[14] Similarly, as shown in Figure 2.3, tweet "*Here's my review of #Arsenal, otherwise known as the first Nicolas Cage film of 2017 http://substreammagazine.com/2017/*

[8] https://tagme.d4science.org/tagme/.
[9] http://www.zemanta.com/.
[10] https://developer.aylien.com/text-api-demo.
[11] https://babelnet.org/.
[12] http://en.wikipedia.org/wiki/Arsenal_F.C.
[13] http://en.wikipedia.org/wiki/Arsene_Wenger.
[14] http://en.wikipedia.org/wiki/Tottenham_Hotspur_F.C.

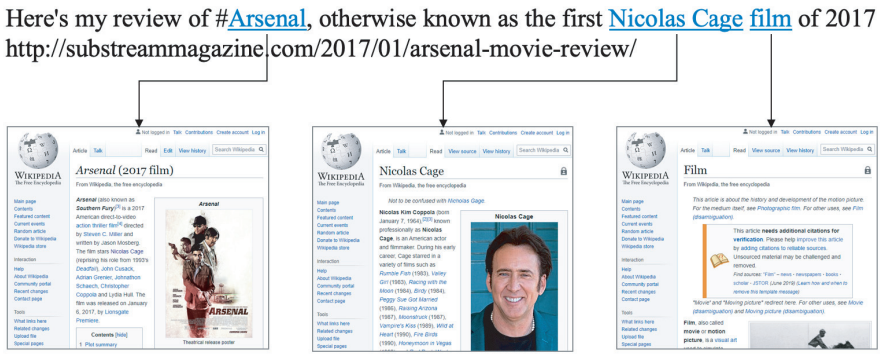

Figure 2.3: Extracted Wikipedia entities for tweet "*Here's my review of #Arsenal, otherwise known as the first Nicolas Cage film of 2017 http://substreammagazine.com/2017/01/arsenal-movie-review/*".

01/arsenal-movie-review/", is annotated with Arsenal_(2017_film),[15] Nicolas_Cage[16] and Film.[17] Therefore, by representing each user's topic of interest by a Wikipedia entity, the users who are engaged with each of these tweets will be identified to have different interests (the first tweet is related to *Arsenal Football Club* while the second tweet is related to *2017 American film*).

Another type of concept for representing user interests is *category*. Using categories, it is possible to represent more general user interests compared to using entities. Four types of categories are mostly utilized in the literature to represent users interests: (1) Wikipedia or DBpedia categories, (2) news categories, (3) Open Directory Project (ODP) categories (known as DMOZ[18] categories), and (4) Pinterest categories.

Most of the studies use Wikipedia categories since the relation between Wikipedia entities and categories are explicitly presented in Wikipedia (Faralli *et al.*, 2017; Kapanipathi *et al.*, 2014; Zarrinkalam *et al.*, 2017). Therefore, it is easily possible to first use existing entity linking tools to annotate the users' textual content to Wikipedia entities and then infer the related categories of interest. It is also possible to

[15] https://en.wikipedia.org/wiki/Arsenal_(2017_film).
[16] https://en.wikipedia.org/wiki/Nicolas_Cage.
[17] https://en.wikipedia.org/wiki/Film.
[18] https://dmoz-odp.org/.

Figure 2.4: Wikipedia categories for entity Arsenal_F.C.

Figure 2.5: Wikipedia categories for entity Arsenal_(2017_film).

extract more abstract categories by following the subcategory links between different categories in the Wikipedia category graph.

Figures 2.4 and 2.5 show the categories that are explicitly mentioned at the end of Wikipedia entities Arsenal_F.C. and Arsenal_(2017_film), respectively. For example, three categories Football_clubs_in_England, Premier_League_clubs and Football_clubs_in_London can be used for representing high-level interests of a user who is interested in Arsenal_F.C. As another example, 2017_films, English-language_films and American_films are three potential categories of interest for a user who is interested in Arsenal_(2017_film).

An example for extracting more abstract categories of interest for a user, by walking through the associated categories of her entity of interest (i.e., Arsenal_F.C.), is shown in Figure 2.6. In this example, we start with the category Premier_League_Clubs from Arsenal_F.C. entity, which produces three categories. Then, we show one step further, by showing the super-categories of category Association_football_clubs_by_competition. The bottom of the figure shows the Wikipedia category graph for further walking through the categories.

In some studies, it is argued that Wikipedia categories tend to get out-of-date and do not keep up with the real time nature of social media such as Twitter. As a result, for example, based on the idea that social media and news media are similar because many current issues are posted in both, some studies first map users' posts to news categories and then represent user interests using news categories

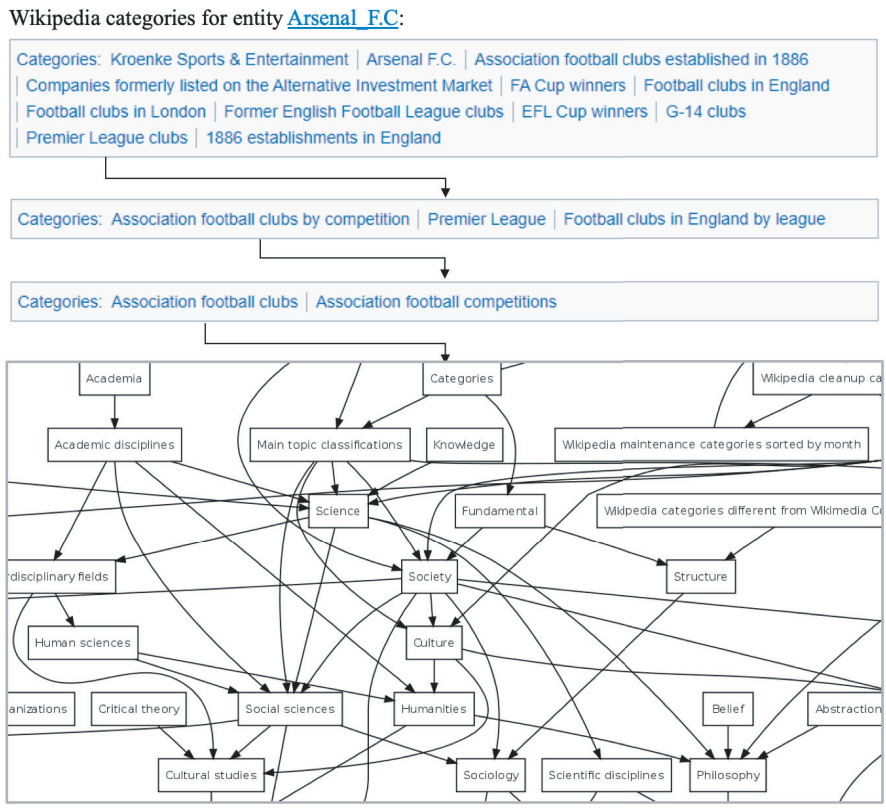

Figure 2.6: Walking through the Wikipedia category graph starting from category Premier_League_Clubs from Arsenal_F.C. Wikipedia article.

(Han and Lee, 2016; Kang *et al.*, 2019). The categories defined in Open Directory Project taxonomy (known as DMOZ categories) are also used in some studies for representing users' interest. It is based on the idea that the categories in this taxonomy provides a clear and broad coverage of various real-world interests and they are much closer to the intent of Twitter usage (Kang *et al.*, 2019). Table 2.2 provides an example of DMOZ main categories and their subcategories that can be used for representing users' categories of interest.

As another type of categories to represent the high-level interests of users, some studies model interest profile of users over the predefined

Table 2.2: DMOZ main categories and their subcategories

Category	Subcategories	Category	Subcategories
Arts	Movies, Television, Music, ...	Shopping	Clothing, Foof, Gifta, ...
Games	Video Games, RPGs, Gambling, ...	Reference	Maps, Education, Libraries, ...
News	Media, Newspapers, Weather, ...	Home	Family, Consumers, Cooking, ...
Regional	US, Canada, UK, Europe, ...	Computers	Internet, Software, Hardware, ...
Society	People, Religion, Issues, ...	Science	Biology, Psychology, Physics, ...
Business	Jobs, Real Estate, Investing, ...	Recreation	Travel, Food, Outdoors, Humor, ...
Health	Fitness, Medicine, Alternative, ...	Sports	Baseball, Soccer, Basketball, ...

categories on Pinterest social media as illustrated in Table 2.3 (Cinar *et al.*, 2015).

Each aforementioned representation model has its strengths and weaknesses. Therefore, instead of using a single interest format (entities or categories), *hybrid* models combine different interest units to represent user interests. For example, in Faralli *et al.* (2017), users' interests are

Table 2.3: The 32 predefined Pinterest categories

Animals	Film music books	Home decor	Quotes
Architecture	Food and drink	Humor	Science nature
Art	Gardening	Illustration posters	Sports
Cars motorcycles	Geek	Kids	Tattoo
Celebrities	Hair beauty	Men's fashion	Technology
Design	Health fitness	Outdoors	Travel
Diy crafts	History	Photography	Weddings
Education	Holidays events	Products	Women's fashion

represented using both Wikipedia entities and categories. Similarly, in Piao and Breslin (2016e), both DBpedia entities and WordNet synsets are utilized for user interest representation.

Using concepts instead of keywords to represent user interests not only address some problems of keyword-based approaches, but also the background knowledge of these concepts can be used to further extend the user interests by considering the relationship defined in the knowledge base between concepts (Mendes *et al.*, 2011; Varga *et al.*, 2014). For example, if we represent users' interests using Wikipedia entities, it is possible to use Wikipedia category hierarchy to extract high level interests of users. By using concepts to represent user interests, it is also possible to utilize the background knowledge of concepts in order to characterize user interests. For example, in Orlandi *et al.* (2013), the specificity of an entity is measured based on the incoming and outgoing links of each entity in DBpedia. Their idea is that very specific entities such as Arsenal_F.C. have many different types of outgoing predicates compared to the incoming ones (Orlandi *et al.*, 2013).

Representing each user's interest with a single concept struggles with inferring more specific topics which are only expressible by combining multiple related concepts. In addition, in this representation model, topics of interest are confined to a set of predefined concepts, and it is not possible to identify emerging interests which are not yet expressed in the knowledge base concepts. For example, in November 2010, Jack Wilshere, England and Arsenal footballer, received a caution for common assault over a street brawl, which received much attention on Twitter. Looking at Wikipedia, there is no entry dedicated to this event. As a result, by considering only the predefined Wikipedia concepts as topics of interest, it is not possible to unambiguously and comprehensively describe this topic (Zarrinkalam *et al.*, 2015).

To address the limitations of considering a single concept to represent each interest of a user, the *group of concepts* representation model represents each topic of interest by a group of concepts which are temporally related together. The group of concepts representation model for representing each topic of interest can be learned via unsupervised approaches in a similar manner for learning topics with topic modeling approaches for the group of keywords representation models.

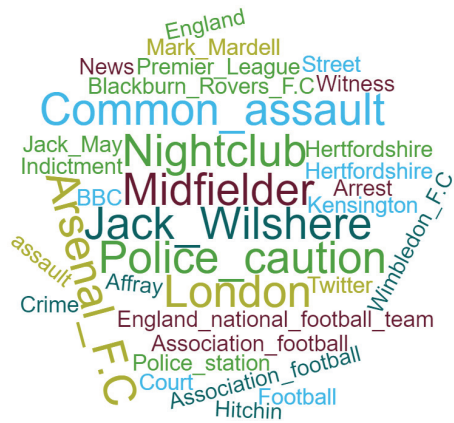

Figure 2.7: A topic that represents a caution received by Jack Wilshere.

For example, as illustrated in Figure 2.7, the conjunction of Wikipedia entities such as Jack_Wilshere, Common_assault, Arsenal_F.C., Nightclub and London can form a topic of interest to represent a caution received by Jack Wilshere for common assault over a street brawl. Therefore, even if a single corresponding semantic concept is not available in the external knowledge base for this event, they are able to construct its semantics by using existing concepts. Furthermore, the group of concepts representation unit not only captures the semantics of the underlying content, but also makes it possible to model more specific interests of users.

2.3 Temporal User Interest Modeling

Based on the fact that users' interests change over time, temporal aspects have been widely used for the conventional recommendations and user modeling in social media (Liang *et al.*, 2018). Figure 2.8 shows interest distribution of two sample users (User A – left; User B – right) over six topics as well as the main words of each topic. For example, User A is highly interested in Dating topic in the first 10 days and her degree of interest in this topic is declining over time. While her interest rate to Celebrity and Health remained stable in the last month, she developed a short-term interest in Baseball. On the other hand, user B

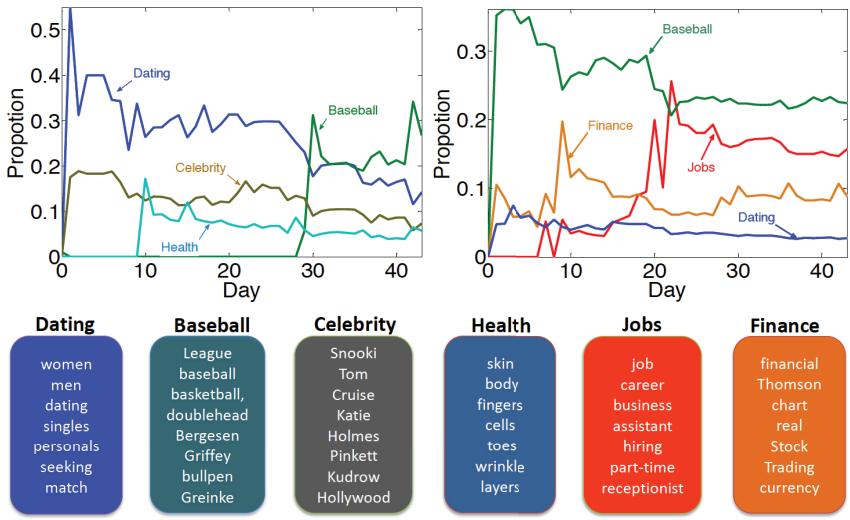

Figure 2.8: Dynamic interest of two sample users (Ahmed *et al.*, 2011).

has a long-term interest in the Baseball topic and her degree of interest to Jobs increased dramatically in the last month.

In order to incorporate dynamicity of user interests into user interest modeling strategies, there are two main types of approaches (Piao and Breslin, 2018a): (1) sliding window, and (2) decay function.

Sliding Window. There are two types of sliding window functions, whose window size is defined by: (1) the number of items (e.g., the recent 100 tweets), and (2) the short period of time (e.g., one week). One basic solution which is followed in most of the studies is to capture users' online behavior over time and build user profiles at different time intervals separately. For example, in the domain of tweet recommendation, in Khater *et al.* (2014), the dynamic level of user interests is captured by building user's interest profiles daily over the extracted topics from Twitter. Then, the history of the user's interest in a window of the last one week is used to calculate the user's degree of interest in a tweet. Similarly, a personalized tweet recommendation system is proposed in Elmongui *et al.* (2015) that creates the dynamic user interest profiles by modeling them as a time-variant in different topics to accommodate the

change of these interests over time. The degree of interest of each user in each topic is calculated over a sliding window of one week. As another example, in Yin *et al.* (2015), the time is divided into multiple time intervals at a predefined granularity and then a unified probabilistic model is proposed that extracts both user-oriented topics as intrinsic interests of users and time-oriented topics as temporal context that attracts public attention.

Decay Function. Instead of constructing user interest profiles in a sliding window, decay-based user interest modeling aims at including all the interests of a user but decaying old ones. Therefore, many researchers have focused on applying time decay functions over historical user generated content of users (i.e., long-term user interest profile). Based on time decay functions, the weight of each interest is calculated based on its age. The intuition behind those interest decay functions is that a higher weight should be given to recent interests than old ones.

For example, in Abel *et al.* (2011a,b), it is shown that a user's interests change over time and are influenced by public trends in twitter. They have modeled user interests in a given timestamp as a set of weighted concepts and have proposed a time-sensitive interest decay function according to the temporal distance between the concept occurrence time and the given timestamp for calculating the weight of each concept. Similarly, in Orlandi *et al.* (2012), the authors have investigated applying an exponential time decay function to compute the user interests weights through a user study. They have shown that using a slower decay function to represent user interests provides a more complete picture of the user and hence produces more accurate profiles. In Ahmed *et al.* (2011), the authors have argued that because of the sparsity of user's history, using exponential decay function would have forgotten quickly about the user's long-term interest, therefore, their temporal model combines three levels of abstractions: (1) user interests over the whole user history, (2) interests of the user in the recent week, and (3) interests of the user in the last month.

The effectiveness of the temporal approach incorporated in different studies to capture the dynamics of user interests depends on the purpose of user modeling. For instance, recently, the effectiveness of different

time decay functions for incorporating dynamics of user interests in the context of personalized link recommendations on Twitter is investigated in Piao and Breslin (2016c). They have shown that using decay functions to build user long-term profiles shows noticeable improvement in the quality of link recommendations compared to user profiles without considering any decay of user interests. However, in the context of publication recommendation, in Nishioka and Scherp (2016), it is shown that short-term user interest modeling outperforms applying exponential decay function for building user interest profiles.

2.4 Semantics-Enabled User Interest Profile Representation

In this section, we introduce different semantic web methodologies and technologies adopted in the literature for representing the constructed user interest profiles. Unfortunately, in the field of user interest modeling from social media, there exists no established industrial standards to support the selection of an ideal representation model. In literature, whether the representation model for individual topics of interest is a keyword-based or a concept-based model, the most adopted model for representing the final user interest profile is the generic Entity-Relationship (ER) model.

Introduced in 1976, an ER model (Chen, 1976) is composed of two main elements: (1) those representing the types of entities in the domain of interest, and (2) those specifying different relationships that can exist between the instances of those entities. ER model representations, due to their simplicity and effective support of database technologies, which are also able to provide a comfortable access to data sources even in the context of big data applications, have been widely adopted in many fields of research and applications.

In the field of user interest modeling from social media, the generality and simplicity of ER representations leads to undesired side effects, mainly affecting the interoperability of the resulting user interest models (Bojars *et al.*, 2008; Orlandi *et al.*, 2012; Peña *et al.*, 2013). To remedy the situation, Semantic Web (SW) technologies, for instance ontologies, have been used in some studies with the purpose of improving interoperability and tackling difficulties of integrating data from multiple sources. This

line of studies has shown promising results, emphasizing the need for further investigations in this direction. For example, SW technologies are involved in Ma *et al.* (2011) to favour the modeling of user interests across multiple social media or as shown in Piao and Breslin (2016c) to model and represent both the dynamics and semantics of user interests.

In the following, we first provide some background information about semantic web technologies and different semantic web vocabularies used in the literature to represent user interest models. Then, in Subsection 2.4.2, we review some sample studies to show how these vocabularies are adopted to represent the constructed user interest profiles.

2.4.1 Semantic Web Technologies and Vocabularies for User Interest Modeling

The term "SW technologies and vocabularies" identifies the ensemble of existing methodologies and technologies developed since 2006, the year when Nigel Shadbolt and Tim Berners-Lee published a foundational article describing the vision of the Semantic Web (Shadbolt *et al.*, 2006). In the SW vision, information is shared on the Web in a form which is understandable by both humans and machines. In fact, the semantic web depicts a scenario where the knowledge representation models are based on a more interoperable framework so that data can be shared and reused across applications. This, among many other interesting benefits, has the potential to overcome the limitations of ER representation models. This enables information to be published and shared on the Web as a global interlinked knowledge graph where both nodes and relations are formally described using a collection of well-defined core ontologies. Further extensions and complementary additions to this core ontological layer can be developed in the form of ontologies which provide domain/application specific knowledge.[19]

From the point of view of the underlying technology stack, different technologies are introduced by Berners-Lee (2005) as the main building blocks of the Semantic Web. Among these, the Resource Description Framework (RDF)[20] provides a data model for formally describing

[19] https://www.w3.org/standards/semanticweb/.
[20] https://www.w3.org/TR/rdf-concepts/.

resources and relationships between them. It is worth noting that the relationships can at the same time be considered as specific resources, and hence their properties and relationships can be described using RDF. An excerpt of RDF specification is composed of RDF statements called triples, where each triple is in the form of a subject-predicate-object expression. The subject and object represent two resources and the predicate represents a specific relationship between those resources. On top of RDF lies the RDF schema which provides the primary constructs, in terms of a set of classes and properties, to describe ontologies.[21] A set of RDF statements can be serialized in different formats such as RDF/XML and N-triples to make the data interchangeable across different applications. Furthermore, RDF data can be stored in RDF repositories and queried using query language such as SPARQL.[22] Finally, the formalism provided due to the use of ontologies, has the benefit that it enables automatic reasoning over ontological representation of data. This reasoning can be realized in the form of automatic detection of inconsistencies in the data, or inferring not-explicitly stated information based on the specification of the relations' characteristics, e.g., being transitive or symmetric.

In the following, different SW vocabularies frequently used to represent user interest models are introduced.

Friend Of A Friend (FOAF) Vocabulary.[23] The Friend Of A Friend (Graves *et al.*, 2007) is a vocabulary which enables the description of social interactions between entities belonging to three kinds of network: (1) social networks of human collaboration, friendship and association, (2) representational networks that describe a simplified view of a cartoon universe in factual terms, and (3) information networks that use Web-based linking to share independently published descriptions of this interconnected world.

As also clearly stated on the FOAF specification, the scheme is not restricted and limited to the description of user relationships on social media, but is able to describe even more general relationships in real

[21] https://www.w3.org/TR/rdf-schema/.
[22] https://www.w3.org/TR/rdf-sparql-query/.
[23] http://xmlns.com/foaf/spec/.

and fictional scenarios. In the context of user interest modeling on social media, the FOAF vocabulary represents a valid scheme to represent user network related relationships. For example, Mypes (Abel *et al.*, 2010), an online service designed to connect, aggregate, align and enrich users' profile information from diverse online services, provides users' profiles expressed on top of the FOAF vocabulary.

Semantically Interlinked Online Community (SIOC) Vocabulary.[24] In contrast to FOAF vocabulary that aims to model users and their networks, Semantically Interlinked Online Communities (SIOC) project (Breslin *et al.*, 2006) provides a lightweight ontology for describing user-generated content in the Social Web and the structure of online communities. The SIOC ontology can be combined with other ontologies such as FOAF and allows interlinking information across online communities. For example, in Bojārs *et al.* (2008) different ontologies such as SIOC and FOAF are used for interlinking and reusing user data across different social applications. With respect to user interest models, SIOC provides a valid representation scheme to model entities such as users (*sioc:User*), posts (*sioc:Post*) and topics (*sioc:has_topic*).

Simple Knowledge Organization System (SKOS) Vocabulary.[25] The Simple Knowledge Organization System (SKOS) (Miles *et al.*, 2005) provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary. Hence, the SKOS vocabulary favors the integration and interlinking of concepts across different knowledge sources. In the context of concept-based user interest models, SKOS properties, such as *skos:related* and *skos:exactMatch*, are used to represent the mapping of user interests into existing entities defined on external knowledge bases.

Social Semantic Cloud of Tags (SCOT) Vocabulary.[26] The SCOT (Social Semantic Cloud of Tags) is an ontology for representing

[24] https://www.w3.org/Submission/sioc-spec/.
[25] http://www.w3.org/2004/02/skos/core.html.
[26] http://rdfs.org/scot/spec/.

tag data and social relations across different sources by different classes such as *scot:Tag* and *scot:Tagcloud*. It provides the structure and semantics for describing resources, tags, and users, and includes some properties such as *scot:acronym*, *scot:spelling_variant* and *scot:synonym* in order to reduce tag ambiguity. For example, in Abel *et al.* (2010), the tag-based profile of a user is extracted from the user's latest tagging activities in Flickr, Delicious, StumbleUpon and Last.fm and represented by SCOT vocabulary.

Weighted Interests (WI) Vocabulary.[27] The Weighted Interests (WI) vocabulary provides a model to describe groups of ordered user preferences. The WI vocabulary enables the specification of relations of interest between instances of the class *foaf:Agent* and instances of the class *wi:WeightedInterest* by the property *wi:preference*. It can also be used to explain that a user (an instance of the class *foaf:Agent*) is not interested in a collection of documents (*foaf:Document*) or in a collection of concepts (*skos:Concept*) by the property *wi:notInterestedIn*. In the more recent revisions of the vocabulary[28] particular attention is posed on interest dynamics. With the class *wi:InterestDynamics*, it is possible to define a specific interval of interest validity, enabling a better representation of the dynamics between users and interest.

There are also some other vocabularies such as **Weighting Ontology (WO)**[29] and **Open Provenance Model (OPM)**[30] which are used in the literature to provide more information about user's interests. The WO is used to relate any topic of interest to a *wo:Weight* instance with the property *wo:weight*. It includes *wo:weight_value* and *wo:scale* properties to associate the numeric value and the scale of the weight respectively. The OPM offers several core concepts and relationships to represent provenance. It includes *opm:wasDerivedFrom*, through which it is possible to model the resource that each topic of interest for a given user is extracted.

[27] http://smiy.sourceforge.net/wi/spec/weightedinterests.html.
[28] http://smiy.sourceforge.net/wi/spec/weightedinterests.html#sec-example.
[29] http://smiy.sourceforge.net/wo/spec/weightingontology.html.
[30] http://openprovenance.org/.

2.4.2 Sample Ontological User Interest Profile Representation

In this section we show some examples of user interest model representations based on SW technologies. For example, GeniUS (Gao *et al.*, 2012) is a software library for generating semantically meaningful social media profiles based on the concepts that are mentioned in the tweets of users. The user interest profiles generated by GeniUS are represented by both FOAF and Weighted Interests (WI) vocabularies. The library also supports a semantic filtering feature through which application developers can specify a SPARQL query that describes what kind of topic-based profile a client is seeking for in a domain specific application.

Figure 2.9 shows the extracted interest profile for a sample user. It includes the user's topics of interest from different domains such as music, software and movies. Using *wi:preference* property, the profile shows that the user is interested in jazz music (*dbpedia:Jazz*), short movies (*dbpedia:Short_film*) and software products (e.g., *dbpedia:Second_Life*). The degree of interest of the user to each topic is specified by Weighting Ontology (i.e., *wo:weight*). The profile depicts that the interest of the user to jazz music is 0.5889 which is more than his interest in short movies which is 0.3333.

Given the RDF repository of interest profile, using the following SPARQL query, GeniUS library will extract only the concepts that belong to the software domain (i.e., the extracted interest profile only includes *dbpedia:Second_Life* and *dbpedia:GarageBand*).

Similarly, in Kapanipathi *et al.* (2011), interest profile of users are represented using FOAF and the Weighted Interests (WI) vocabularies. A sample interest profile is depicted in Figure 2.10. Since they have inferred the users' interests from multiple social media, for each topic of interest, they have specified its source using Open Provenance Model (OPM)[31] vocabulary. Based on Figure 2.10, the user is interested in Semantic Web which is derived from (*opm:wasDerivedFrom*) both Twitter and LinkedIn.

As other examples, in Kapanipathi *et al.* (2011), the authors have presented an approach to suggest public tweets according to a user's interests. Since, they have represented both users' profiles and tweets in

[31] OPM Specification: http://openprovenance.org/.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix wi: <http://purl.org/ontology/wi/core#> .
@prefix wo: <http://purl.org/ontology/wo/core#> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix genius: <http://persweb.org/genius#> .
<http://twitter.com/bob>
  a foaf:Person;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:Jazz ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.5889 ;
        wo:scale genius:Scale ]
     ] ;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:Second_Life ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.3114 ;
        wo:scale genius:Scale ]
     ] ;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:Short_film ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.3333 ;
        wo:scale genius:Scale ]
     ] ;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:GarageBand ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.1638 ;
        wo:scale genius:Scale ]
     ] ;  ...
```

Figure 2.9: Example of the extracted interest profile for a sample user (Gao *et al.*, 2012).

RDF, thanks to semantic web technologies (i.e., RDF triples stores and SPARQL), the presented approach is able, in real time, to semantically select public tweets from public streams and suggest those semantically related to the interests of a specific user.

Recently, in Tommaso *et al.* (2018), a multi-domain interest dataset named Wiki-MID is collected and published for Twitter users which is published based on SW technologies and vocabularies. As shown in Figure 2.11, in Wiki-MID, Twitter users are instances of the *sioc:User-Account* class. The relation between each user and her interests extracted from her posts are specified by *sioc:likes* predicate; and *skos:relatedMatch* predicate is used to show the corresponding Wikipedia entity of each user's interest which is the result of the automated mapping methodology applied to map the user's interests to Wikipedia entities.

```
<foaf:topic_interest rdf:resource="http://dbpedia.org/resource/Semantic_Web" />
<wi:preference>
  <wi:WeightedInterest>
    <wi:topic rdf:resource="http://dbpedia.org/resource/Semantic_Web" />
    <rdfs:label>Semantic Web</rdfs:label>
    <wo:weight>
      <wo:Weight>
  <wo:weight_value rdf:datatype="http://www.w3.org/2001/XMLSchema#double">0.5</wo:
      weight_value>
  <wo:scale rdf:resource="http://example.org/01Scale" />
      </wo:Weight>
    </wo:weight>
    <opm:wasDerivedFrom rdf:resource="http://www.twitter.com/BadmotorF" />
    <opm:wasDerivedFrom rdf:resource="http://www.linkedin.com/in/fabriziorlandi" />
  </wi:WeightedInterest>
</wi:preference>
[...]
<wo:Scale rdf:about="http://example.org/01Scale">
  <wo:max_weight rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">1.0</wo:
      max_weight>
  <wo:min_weight rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">0.0</wo:
      min_weight>
</wo:Scale>
```

Figure 2.10: Representing an interest and its weight found in two sources (Kapanipathi *et al.*, 2011).

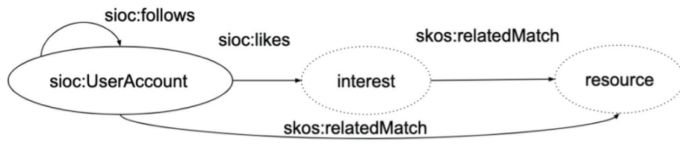

Figure 2.11: The model representation of Wiki-MID dataset (Tommaso *et al.*, 2018).

2.5 Cross-System User Interest Modeling

User interest modeling from social media can be viewed as either *single-system* or *cross-system* (Orlandi *et al.*, 2012). In single-system approaches, only one social media is considered as a source of information for user interest modeling. Cross-system approaches, on the other hand, are based on the idea that a user has different profiles in different social media for different purposes, and to extract her interests more accurately, it would be better to extract and integrate her information from all those profiles (Spasojevic *et al.*, 2014).

For example, users on LinkedIn connect to their business partners, Facebook users connect to their friends, Twitter users probably post about recent news and events, tag-based Flickr profiles are related to geographical topics, while Delicious and StumbleUpon profiles refer to topics in the area of communication.

More specifically, in Ottoni *et al.* (2014), the authors have analysed the activities of users who have accounts on both Twitter and Pinterest and observed that although many users seem to maintain a single identity and interests across the two social media, they do different activities on each social media. For example, by modeling user's interests on each social media over Pinterest's predefined categories, they observed that users tend to engage in more categories on Pinterest than on Twitter, and, while Twitter is a popular communication platform, Pinterest plays a key role in identifying new interests of users. Therefore, since the overlap of the individual user profiles across the different social media is rather low (e.g., on average, less than 10% for Flickr and Delicious profiles), by aggregating the information about a user from multiple social media, it is possible to build a *multi-faceted* user profile which is more comprehensive compared to single-system profiles (Abel *et al.*, 2013b).

Another advantage of cross-system user interest modeling is that they can improve the quality of recommender systems especially by addressing the *cold-start* and *sparsity* problem (Abel *et al.*, 2013b). For example, in Abel *et al.* (2013b), the authors have built a tag-based profile of a user by aggregating the user profiles extracted from the social tagging activities of the user in different social media such as Flickr, Delicious and StumbleUpon. By evaluating the performance of cross-system user modeling strategies in the context of tag recommendation systems, they have shown that the aggregated profile is able to solve the cold-start problem and improve the quality of recommendations, even beyond the cold-start.

As another example, based on the idea that a user might be more active on Google+[32] than Twitter, in Piao and Breslin (2016a) the information of a user on Google+ is utilized to improve her recommendations on Twitter. Further, despite most of the cross-system user modeling studies that have aggregated the inferred interest profile from each social media with the same weight (e.g., Abel *et al.*, 2013b), they have applied different weights to each social media profile for the aggregation process.

[32]Google+ no longer available for consumer and brand accounts.

Although cross-system user modeling brings many opportunities, taking advantage of these opportunities is not trivial. The first and foremost step of cross-system modeling is user *identity linkage* across online social media (Shu *et al.*, 2016). For example, in Abel *et al.* (2013b), the accounts of a given user in different social media (e.g., Facebook, Twitter and LinkedIn) are identified by exploiting the *Google Social Graph API*, that provides account mappings for all users who linked their accounts via their Google profile. Similarly, in Piao and Breslin (2016a), the accounts of users who use both Twitter and Google+ are obtained from about.me[33] dataset.

As user identity linkage across social media is a very important and challenging problem, it has become a trending research area and attracted more and more research attention. The main challenges of user identity linkage are as follows (Shu *et al.*, 2016): (1) Profile Inconsistency: Different social media have different structures to present user profile attributes, e.g., age, location, friends. Further, the same attribute can be filled up with different information by a user in different social media based on the user's purpose in that social media. Therefore, profile inconsistency increases the uncertainty and ambiguity mapping the user's accounts across different social media. (2) Content Heterogeneity: The user's generated content may be in various medium types such as text, image, video and check-ins in different social media. Therefore, The heterogeneous content information makes it difficult to use them to measure the similarity between the users' behavior accurately to link user identities. (3) Network Diversity: Users may be connected to different users in different social media based on the user's purpose in that social media. For example, a user may be connected with her co-workers on LinkedIn who are not necessarily the same as her friends on Facebook. This may prevent us from using graph structure patterns to match user entities as traditional entity resolution tasks. Interested readers about the state-of-the-art approaches on user identity linkage across social media are encouraged to see (Shu *et al.*, 2016).

[33] https://about.me.

3

User Interest Modeling Approaches

Most of the studies in user interest modeling from social media extract users' *explicit* interests that are directly observable from user content. However, given the increasingly noticeable *free-rider* behavior in social media, techniques that rely on users' explicitly observed content may not be fully desirable in such circumstances. Some other techniques focus on passive users and extract their *implicit* interests by considering the interaction patterns between users and topics. There is another line of work that is dedicated to predicting users' *future* interests instead of modeling current or past interests of users. These studies are primarily focused on predicting if and which users would be interested in future topics on social media.

In this section, we introduce different approaches proposed in each category, i.e., *explicit user interest detection*, *implicit user interest mining* and *future user interest prediction*.

3.1 Explicit User Interest Detection

There is a rich line of research on user interest detection from social media that have focused on extracting *explicit* interest of a user by leveraging information from the user's own activities (e.g., social posts

and social relations). The intuition behind explicit interest detection is "You are what you share". Take Twitter as an example, a user can have different activities such as posting, re-tweeting, liking or replying to a tweet or follow other users. For example, if a user mentions the term Arsenal frequently in her tweets or follows the Twitter account @Arsenal, one can conclude that she is explicitly interested in Arsenal football club. Similarly, if a user frequently posts some photos related to football on social media platforms such as Flickr, one can conclude that she is interested in football.

Given a set of topics of interest (represented by keywords or concepts), based on the weighting scheme that is utilized to determine the importance of the interest with respect to a user, the explicit interest detection approaches can be divided into three categories: (1) *frequency-based* approaches, (2) *probabilistic* approaches, and (3) *machine learning-based* approaches, and (4) *similarity-based* approaches.

3.1.1 Frequency-Based Approaches

A common and simple weighting scheme to calculate the degree of interest of a user to a given topic of interest is using the frequency of the interest (e.g., keyword or concept) in the user's activities. For example, to detect explicit interests of uses, the authors in Abel *et al.* (2011c) first enriched the Twitter posts by linking them to related news articles in different news media such as CNN, BBC or New York Times. Then, for a given user, they have extracted the entities (e.g., people, organization or events) or topics (e.g., politics and sports) of the user's enriched posts using web services provided by OpenCalais.[1] Finally, to build a user interest profile (entity-based or topic-based profile), they have calculated the degree of interest of the user to each entity/topic by counting the number of her tweets referring to the given entity/topic. Similarly, in Nguyen *et al.* (2016), to build a user interest profile, a multi-label classifier is trained to classify each Twitter post into one or several topics out of 16 label genres (15 from Francis and Kucera (1979) and an additional label named "tech"). Then, the interest profile of a user is constructed by aggregating all the occurrences of each genre

[1] https://developers.refinitiv.com/open-permid/intelligent-tagging-restful-api.

for each post written by the user. Finally, the constructed user interest profiles have been used for recommending articles from another social media platform – Reddit,[2] which is a network of communities based on people's interests.

In Dinh and Van Pham (2020) the authors have proposed a CNN-based deep learning model to classify Facebook posts into five topics such as "Health" and "Technology" by considering word embeddings with user features such as gender, age and the number of followers. Afterwards, a user interest profile is constructed based on the number of her posts tagged by each topic. The similar idea of using multi-label classifier(s) for inferring user interest profiles has been applied for photos as well. The authors in Pandey and Sang (2015) have trained two-level classifiers to classify photos shared by users on Instagram into topics in an interest ontology with two levels. Then, a user interest profile is built based on the percentage of her photos belonging to each of the interest topics/categories.

To model explicit interests of users over DBpedia entities, in Orlandi *et al.* (2012), Zemanta, an entity annotator, is used for connecting a user's textual content to DBpedia entities. To build an entity-based profile for the user, the degree of interest of the user to each extracted entity is calculated by applying an exponential time decay function to the frequency value of the entity in the user's textual content to give higher weight to the interests that occurred recently. Similarly, in Kapanipathi *et al.* (2011), the authors modeled the users' interests by annotating their social posts with DBpedia entities. However, instead of using an existing entity annotator, they applied a dictionary-based technique in which the dictionary includes a set of concepts from DBpedia. In their model, the degree of interest of a Twitter user to an entity is calculated by dividing the number of occurrences of the entity in her latest 500 posts by the total number of entities identified in the same posts.

Photos shared by users have been also used for building concept-based user interest profiles where the vocabulary is a set of predefined labels obtained by classifying those photos using a classifier such as

[2] https://www.reddit.com/.

GoogleNet (Szegedy *et al.*, 2015) or deep learning based image annotation frameworks such as Caffe (Jia *et al.*, 2014). For example, the problem of extracting interests from shared photos is treated as an object classification task in Wieczorek *et al.* (2018). The authors have adopted existing computer vision approaches such as ResNet-50 to classify each photo into 1000 ImageNet concepts. For each classified label, the authors have used the semantic network of BabelNet to find a domain for each label using predefined rules. Finally, a user interest profile is built based on the frequency of each domain (34 in total) extracted from shared photos on Flickr.

The authors in Wang *et al.* (2009) have used an image annotation tool, named Arista (Wang *et al.*, 2008), to annotate each photo shared on Flickr by users. Given each category in DMOZ taxonomy as a topic, they have measured the degree of interest of a user to each topic using the frequency of the words in those annotations in topics. As a naive model, the authors have also utilized tags assigned to photos to build a tag-based interest profile. Similarly, in Joshi *et al.* (2015) the interest profile of a user is built based on analysing the visual content of photos shared by the user on Flickr where the vocabulary is the 1000 ImageNet concepts. Given the concepts and scores assigned to each photo using the Caffe image annotator, for each user, her interest in a concept is calculated by the average of the concept scores across her photos. The authors also built a tag-based interest profile for a user based on the tags assigned to each photo by its owner or third party users and the weighting scheme is a standard TF-IDF based score vector across the vocabulary of tags.

Instead of representing each topic of interest as a single concept, in Zarrinkalam *et al.* (2015), the authors have viewed each topic of interest as a conjunction of several concepts and modeled user explicit interests over these topics by analysing the user's textual content. To extract active topics, they have first constructed a concept graph in which nodes are the Wikipedia entities mentioned in the posts (TagMe annotator is used to annotate posts) and the links between each two entities are based on their co-occurrence in posts. Then they have applied a clustering algorithm to extract the group of concepts which are temporally correlated on Twitter as active topics in a given time

interval. Finally, they have calculated the degree of interest of a user to each extracted topic based on the frequencies of the constituent entities of the topic in the posts published by the user.

3.1.2 Probabilistic Approaches

In some studies, the probabilistic topic models such as LDA are utilized to weigh the interests of a user. For example, in Weng *et al.* (2010), the topic profile of each user is extracted by first aggregating the posts published by an individual user in Twitter into a single document and then applying LDA over the collection of documents. The authors have utilized the discovered topic profiles for users to calculate the topical similarity between users in order to measure the influence of users on Twitter. In He *et al.* (2014), a novel topic model called User-Topic model (UTM) is proposed to discover user interest in microblogs, considering both original interest and retweet interest of users. Similarly, in Xu *et al.* (2011) topic modeling approach is utilized for user interest modeling. The authors have indicated that the large amount of posts on Twitter (e.g., social-chatting and daily-life posts) are interest-unrelated. Therefore, they have modified author-topic model (a generative model that extends LDA) (Rosen-Zvi *et al.*, 2004) for extracting the users' interests where the latent variables are used to indicate whether the post is related to the user's interests or not which result in filtering out the interest-unrelated posts from the process of user interest modeling. The author-topic model is also modified in Xu *et al.* (2012), by assuming that user behavior is mainly influenced by three factors: breaking news, posts from social friends and user's intrinsic interest.

Temporal user modeling, which incorporates the timestamp of user behavior data and understands users' interest evolution, has also been considered in Probabilistic approaches to infer explicit interests of users. For example, in Yin *et al.* (2015), based on the intuition that users' behaviors are influenced by user intrinsic interests and the public's attention during a time period, the authors have proposed a temporal context-aware mixture model (TCAM) that explicitly introduces two types of latent topics to model user interests. Since TCAM assumes that the users' interests are stable over time, they have also extended their

model to a dynamic temporal context-aware mixture model (DTCAM) to explicitly model the dynamics of users' interests. The authors have evaluated their model on different social media, i.e., Digg, Douban Movie, MovieLens and Delicious. Similarly, a probabilistic framework is proposed in Sang *et al.* (2015) for inferring user interest profiles. They have assumed users have long- and short-term interest (topic) distributions. Long-term interests denote stable preferences of users while short-term interests denote user preferences over short-term topics of events in social media.

Probabilistic approaches have been also applied to extract topics from both photos and their associated tags. In Xu *et al.* (2011), the authors have proposed a mixture model that leverages both the textual and visual content associated with social images on Flickr. Similar to the idea that a tweet is generated by topics distributions learnt from LDA, the authors assumed a (photo, tags) pair is also generated by topics learned by the proposed approach. To this end, GoogleLeNet (Szegedy *et al.*, 2015) and LDA are applied to extract features from a photo and associated tags respectively. Afterwards, a Gaussian model is used to facilitate the generation from those two spaces.

3.1.3 Machine Learning (ML)-Based Approaches

To weigh the interests of a user, ML-based approaches have been used as well. For example, the authors in Zhao *et al.* (2015) have proposed using matrix factorization techniques to infer a user's topical interests using their actions within Google+. They have represented topical interests as entities in Google Knowledge Graph[3] such as "basketball" or "video games". Given the observation that social media users interact with different topics using different types of behaviors such as commenting, posting, or +1 (liking), the authors introduced a behavior factorization approach considering different behavior types differently (i.e., building matrices of different behavior types) while applying matrix factorization techniques.

There have been some studies focusing on neural network-based embedding methods to learn user interests. For example, in Liang

[3]https://developers.google.com/knowledge-graph.

et al. (2018) the authors have presented a dynamic embedding framework in which words and users are jointly modeled in the same semantic space, which allows measuring the similarity between users and words when constructing a user profile. As other examples, In You *et al.* (2016) the authors have proposed inferring users' interests distribution over 34 topical interests/categories on Pinterest. To this end, the authors first used a Convolutional Neural Network (CNN) to classify each image that has been shared by a user. Afterwards, based on the initial labels predicted by the CNN model for each image, they predict the users' interests from the unorganized collection of images by using a proposed image and group level label propagation approach.

Similarly, the authors in Geng *et al.* (2014) have also studied inferring user interests on Pinterest based on photos shared by users with respect to the fashion domain. Instead of inferring user interests in a set of topical categories, they have inferred ontological models to generate comprehensive user profiles. To this end, they have proposed an ontology construction method to structuralize the curated images onto an ontology by pruning the Wikipedia category graph. Afterwards, a novel multi-task convolutional neural network (mtCNN) is proposed to predict whether an image belongs to a concept in the ontology. Finally, the inferred profiles are further refined by exploiting the rich social relations offered by social curations on Pinterest such as how many users have pinned two images, how many bundles share the two images, and the similarity between those images. Recently, Pinterest also introduced their ML-based approach called User2Interest for inferring user interests. One of the most important input signals to the ML system is user engaged Pins and corresponding interest labels of those pins output from Pin2Interest, which is another ML system that maps Pinterest Pins to interest categories in a predefined interest taxonomy.[4]

To detect explicit interests of users over the categories defined in Reddit, In Fiallos and Jimenes (2019), first a multi-label classifier has been trained where the training dataset was obtained from selected user forums belong to Reddit which had information of users conversations

[4] https://medium.com/@Pinterest_Engineering/interest-taxonomy-a-knowledge\-graph-management-system-for-content-understanding-at-pinterest-a6ae75c203fd.

grouped in eight categories such as sports, entertainment, politics, business and tech. Then, an LDA topic modeling applied over tweets to identify the topics of interest of each user. Finally, the multi-label classifier has been applied to the word topics to model user interest over the eight Reddit categories. In Xu and Lu (2015), the authors have also treated the problem of detecting user interests on Tumblr as a multi-label classification problem. They have proposed a graph model which contains two subgraphs, one for users and the other corresponds to topics of interests (i.e., tags in Tumblr posts). Finally, the user interest detection problem has been formulated as a multi-label classification on the bi-relational graph.

3.1.4 Similarity-Based Approaches

Apart from aforementioned three approaches, some approaches have used similarity/distance measures to derive the weights of each predefined user interest. For example, in Gasparetti (2017), to model the users' explicit interests over eight news categories (e.g., sports, food and politics), the authors mapped both users' posts and news categories to a vector of vocabulary. Given the vectors for the user's posts and news categories, they have calculated the importance of each news category for a given user by simply calculating the cosine similarity between their corresponding vectors. Since there is a semantic gap between the news media and social media, instead of using terms as the vocabulary, they have used Wikipedia entities and categories. For example, social media users often use the term "SNS" instead of the full name "social networking service", while news media use the term "social media service". To solve the semantic gap problem, they have designed Wikipedia-based feature generators to convert a term vector to a vector of Wikipedia categories or a vector of Wikipedia entities. For instance, if a term vector contains the term "Pizza" and another term vector contains the term "Omelette", then their corresponding Wikipedia-based feature vectors may contain the same Wikipedia category (i.e., "Category: food").

Recently, in Seghouani *et al.* (2018), an unsupervised multilingual approach named FRISK (Find twitteR InterestS via wiKipedia) is proposed for extracting a user's interests over a predefined set of Wikipedia

categories (i.e., Politics, Economy, Games, Gastronomy, Sports and Tourism) based on the meaning of the words explicitly mentioned in the user's posts. To build the interest profile of a user, they have first annotated the user's posts to Wikipedia entities using TagMe. Then, given the Wikipedia graph, they calculated the degree of interest of the user to each predefined category based on the graph distance between Wikipedia entities mentioned in the user's posts and the given category representing the interest.

3.1.5 Summary and Discussion

In Subsection 3.1 we reviewed existing approaches in the literature for detecting explicit interests of users by analysing their activities on social media. On one hand, high-quality interest profiles would be expected as those approaches directly process the content that is generated by users to infer their user interest profiles. On the other hand, they require users to be actively generating content on social media platforms, which might not always be the case for a great number of users as we will discuss in the next section. Tables 3.1 and 3.2 summarizes these studies in terms of the basic user interest modeling features introduced in Section 2. In summary, our key findings include:

- The most common and dominant approach for identifying the degree of a user's toward a specific topic is the frequency-based approach through which the frequency of the topic is calculated in the user's activities. However, more recent studies show some other techniques like embedding methods, vectorization and graph analysis can be applied properly for this purpose.

- Social posts (as internal data) are the main source of information for modeling explicit interests of users. Specifically, 28 out of 37 studies (76%) covered in Tables 3.1 and 3.2, use social posts as one of their internal sources of information. However, other sources of information, like images have been used relatively less frequently in the studies. This can be attributed to the fact that most studies are focused on using Twitter as the target social media, which is

Table 3.1: Summary of features (Part 1) used for explicit interest detection from social media

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
			Frequency-Based				
Wang *et al.* (2009)	Images, Tags	DMOZ Taxonomy		DMOZ Categories			
Abel *et al.* (2010)	Posts, Tags	WordNet, Google Social Graph	Tags			FOAF, SCOT	Twitter, Blogspot, Flickr, Delicious, Stumble Upon
Abel *et al.* (2011c)	Posts, URLs	News Media	Named Entities	OpenCalais Topics			
Abel *et al.* (2011b)	Posts, URLs, Tags	News Media	Named Entities, Tags	OpenCalais Topic	Sliding Window		
Kapanipathi *et al.* (2011)	Posts	DBpedia		DBpedia Entities	Sliding Window	FOAF, WI, WO	Twitter, Facebook, LinkedIn
Gao *et al.* (2012)	Posts	DBpedia		DBpedia Entities	Sliding Window, Decay Function	FOAF, WI, WO	Twitter, Facebook, LinkedIn
Abel *et al.* (2011a)	Posts, Tags		Group of Keywords (Named Entities, Tag)		Decay Function		
Yang *et al.* (2012)	Posts, Tags		Tags		Sliding Window		

Continued.

Table 3.1: Continued

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
			Frequency-based				
Abel *et al.* (2013b)	Posts, Tags	WordNet, Google Social Graph, DBpedia	Tags			FOAF, SCOT	Twitter, Blogspot, Flickr, Delicious, Stumble Upon
Joshi *et al.* (2015)	Tags, Images		Tags	ImageNet Concepts			
Cinar *et al.* (2015)	Posts, Images			Pinterest Categories			
Zarrinkalam *et al.* (2015)	Posts	Wikipedia		Group of Keywords (Wikipedia Entities)	Sliding Window		
Elmongui *et al.* (2015)	Posts, Tags, URLs	DMOZ Taxonomy, DBpedia		DMOZ Categories	Sliding Window		
Nishioka and Scherp (2016)	Posts	DBpedia	Group of Keywords	DBpedia Entities	Sliding Window, Decay Function		

Continued.

Table 3.1: Continued

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
			Frequency-based				
Tommaso *et al.* (2018)	Posts, Follow Relation, URLs, Tags	Wikipedia, Spotify, IMDB, Goodreads		Wikipedia Entities		SIOC, SKOS	
Wieczorek *et al.* (2018)	Image	BabelNet		BabelNet Domain Categories			
Nguyen *et al.* (2016)	Posts			Brown Corpus Topic			
Dinh and Van Pham (2020)	Posts			Facebook Page Topic			
Pandey and Sang (2015)	Image	Description about similar images in Google Image Search results		Topic in self-defined taxonomy			

Table 3.2: Summary of features (Part 2) used for explicit interest detection from social media

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
			Probabilistic				
He *et al.* (2014)	Posts		Groups of Keywords (Unigrams)				
Xu *et al.* (2011)	Posts	Wikipedia		Group of Concepts (Wikipedia entities)			
Ahmed *et al.* (2011)	Posts		Group of Keywords (Unigrams)		Sliding Window, Decay Function		
Ottoni *et al.* (2014)	Posts, Tags, Pins			Pinterest Categories	Sliding Window		Twitter, Pinterest
Khater *et al.* (2014)	Posts		Group of Keywords (Unigrams)		Sliding Window		
Yao *et al.* (2017)	Tags, Images	WordNet	Tag				
Yin *et al.* (2015)	Posts, Tags		Group of Keywords (Unigrams)		Sliding Window		
Sang *et al.* (2015)	Posts		Group of Keywords (Unigrams)		Decay Function		

Continued.

Table 3.2: Continued

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
			ML-based				
Zhao *et al.* (2015)	Posts			Google Knowledge Graph Entities			
Cao *et al.* (2017)	Tags		Tag				
You *et al.* (2016)	Images			Pinterest Categories			
Geng *et al.* (2014)	Images	Wikipedia		Wikipedia entities			
Fiallos and Jimenes (2019)	Posts			Reddit Topic			
Xu and Lu (2015)	Tags		Tag				
			Similarity-based				
Kang and Lee (2017)	Posts	Wikipedia, News Media		News Categories			
Liang *et al.* (2018)	Posts		Unigrams		Sliding Window		
Seghouani *et al.* (2018)	Posts	Wikipedia		Wikipedia Categories			

mainly text-centric, hence less potential has been considered to exist for processing images shared over this social media.

- Due to the inherent conciseness of user generated content in the dominant social media, like Twitter, most studies on explicit interest detection utilize an external knowledge source for enriching user contents. Wikipedia and DBpedia are among the main sources that have been used in this regard. This is not surprising, since both sources provide a community-based representation of encyclopedic knowledge which has two interesting properties, i.e., being user-generated and being not limited to a specific domain. Further, due to the availability of high-quality entity annotation tools, like TagMe and Zemanta, that annotate a text with Wikipedia/DBpedia entities, using these knowledge sources for content annotation has become a de-facto standard in the field of user interest modeling from social media.

- Although concept-based techniques have shown to be able to address the limitations of keyword-based techniques for the representation of the user interests, both types of techniques have been and are popular in the studies on explicit interest detection. However, an important observation is that the keyword-based techniques have shifted the interpretation of keywords from unigrams to more information-conveying and larger units like hashtags and named entities.

- While it is well accepted in the literature that users' interests change over time, this dynamicity has not been considered in about half of the studies in explicit interest detection. Among those studies that take temporal aspects of users' interests into account, two groups of works can be distinguished. In the first group, this is considered by simply dividing the time dimension into a set of intervals and extracting the users' interests separately for each time interval based on the interval specific network representation. The second group of studies however deals with the dynamicity of users' interests in a cross-interval fashion, by paying attention to how the user's interests evolve from one interval to the next

one. This evolution is usually modeled using decay functions to simulate the idea that users generally lose interest in the topics in which they have been previously interested.

- Using ontology-based techniques cannot be considered as a noticeable trend among the explicit interest detection approaches. Furthermore, from the analysis of the studies that use ontologies it can be concluded that a very limited number of ontologies have been used in this field. This is not surprising since the potentials of ontologies are realized mainly where there is a tendency towards sharing information in an application-independent way, for use by further parties. Since sharing the detected users' interests has not been the purpose of the studies, it is clear that using ontologies cannot provide noticeable contribution to the field, except for cross-system approaches which deal with detecting users' interests by analysis of her information from multiple social media.

- Despite the interesting potentials of cross-system modeling in providing a more comprehensive and diverse model of the user, due to the existing challenges in user identity linkage over different social media, there is much room for research in order to realize this potential. Most of the existing studies only extract explicit interests of users based on the users' data in one social media and only a limited number of studies propose a cross-system user modeling approach.

3.2 Implicit User Interest Mining

User implicit interests are those potential interests that a user did not explicitly mention but might have an interest in. Since explicit user interest modeling approaches extract interests of a user by using the information from her own activities, it can be only helpful for modeling the interests of active users. However, most users on social media are passive and prefer to passively read rather than to actively engage (Romero *et al.*, 2011). For example, 40% of users browse Facebook

only passively, without posting anything[5] or 44% of Twitter users have never sent a post.[6] Thus, the interests of such users cannot be directly identified from their explicit contributions to the social network. Implicit interest user modeling, not only improves user interest modeling for active users, but also provides the opportunity of modeling users' interests for *free-riders* and *passive users* which are the majority of social media users.

For inferring implicit interests of users, *relationship between users* and *relationship between topics of interest* are two main indicators used in different studies (Zarrinkalam *et al.*, 2016). Based on these indicators, in this section, we review different approaches for implicit user interest modeling which are divided into three categories: (1) *User-centric* models that mainly use the inter-user relationships to infer implicit interests of users, (2) *Topic-centric* models that use inter-topic relationships, and (3) *Hybrid* models that use both indicators in a model.

3.2.1 User-Centric Models

Some authors have shown interest in inter-user relationship information to extract the users' implicit interests. They have usually used the homophily principle (McPherson *et al.*, 2001), which refers to the tendency of users to connect to other users with common interests or preferences, to infer interests of a user based on the information provided by her friends. It is noted that, friends of a user are those who are in the social network of the target user, constructed either based on the explicit links between users on social media or their inferred similarity. For example, based on the homophily principle, in McPherson *et al.* (2001), it is shown that the Twitter users tend to be in contact with other users who are interested in similar topics by representing each user by the hashtags mentioned in her posts.

Textual content of a user's friends is the most common source of information utilized in literature for inferring implicit interests of the

[5]Mander, Jason. "4 in 10 Facebookers Now Browsing the Site Passively." February 20, 2015, https://blog.globalwebindex.com/chart-of-the-day/4-in-10-facebookers-now-browsing-the-site-passively/.

[6]Yoree Koh, "Report: 44% of Twitter Accounts Have Never Sent a Tweet" Wall Street Journal, 11 April 2014, http://bit.ly/2HdpBF2.

user (Wang *et al.*, 2013, 2014). For example, to infer implicit interests of a user from Twitter, a probabilistic model is proposed in Budak *et al.* (2014) which is based on the textual content of the user and her friends. In this model, the authors have incorporated how susceptible the user is to the influence of her friends. Similarly, in Pennacchiotti *et al.* (2012) the interests of a Twitter user are inferred by using the posts published by the user herself and her most authoritative friends. They have characterized authoritative users by a high number of followers and a high ratio between their followers and followees.

In Wang *et al.* (2014), it is argued that the algorithms that are only based on explicit links between the users are too simple and would not be effective in the context of social media which are sparse and dynamic. Therefore, the authors in Wang *et al.* (2014) have extracted user interests based on a specific link structure assumption under which local link structures between two users are considered to be an indicator of user similarity. For example, if two users share many followers, they are likely to be similar in terms of their topical interests. They have first modeled the interest profile of each user separately by analyzing her published posts. Then based on a regularization framework, they have extended the interest profile of a user using the interest profiles of her friends.

More recently, the authors in Gong *et al.* (2020), inspired by the concept of user schema in social psychology,[7] take a new perspective to perform user representation learning by constructing a shared latent space to capture the dependency among different modalities of user-generated data. In their work, both users and topics are embedded in the same space to encode users' social connections and text content, to facilitate joint modeling of different modalities, via a probabilistic generative framework. In some recent studies, it is argued that processing the textual content of a user's friends has some drawbacks: (1) it increases noise in the process of user interest modeling. Because, the friends of a user can publish a wide range of topics that they are interested in, and the user is not always interested in all those topics.

[7] A user schema describes a pattern of thought or behavior that organizes categories of information and the relationships among them (DiMaggio, 2003).

(2) It increases the difficulties of large-scale text processing for user interest modeling (Bhattacharya *et al.*, 2014).

To infer implicit interests of users at scale, *tags*, *biographies* and *list memberships* of the users' friends are other sources of information that have been utilized in some studies. For example, in Cao *et al.* (2017), the authors have defined the problem of learning user topical profiles as inferring the unknown tags of users based on other user topical profiles where their interests/tags are available. Here, the tags are the names of Twitter lists that a user has been added into. As most of target users do not have any partially explicit footprint (i.e., tags), the authors have proposed a matrix factorization based approach, which integrates many different kinds of implicit footprints. Implicit footprints include social footprints such as whether two users have common connections, interest footprints that consider whether two users have common topics in their shared posts, and behavior footprints that consider whether two users have similar behaviors such as retweeting the same tweet or sharing the same URL.

As other examples, in Bhattacharya *et al.* (2014) and He *et al.* (2015b), the implicit interests of a user are extracted based on the topical expertise of her Twitter friends. Their approach is based on the idea that a user might be implicitly interested in a given topic if she is following the users who have been added into many topical lists related to that topic. In He *et al.* (2015, 2020), the authors have identified user's interests for non-famous users based on the expertise of their famous friends. The users who are followed by less than 2000 users are called non-famous users and the users who are followed by at least 2000 users are called famous users. Their approach includes two main components: in the first components, they have extracted the topical expertise of famous users based on their topical lists and in the second component they have proposed a Bi-Labeled LDA to extract interest tags for a non-famous user. Similarly, the authors in Grbovic *et al.* (2016b) have used the signals of *following* and *liking* the posts created by social influencers with respect to specific topics of interest for inferring implicit interests of users in those topics on Tumblr.

To build the interest profile of a user based on the expertise of her friends, despite the approach in Bhattacharya *et al.* (2014), the authors

in He *et al.* (2015b, 2020) have not considered all the expertise of a user's friends as her implicit interests. They have argued that a famous user might be expert in different aspects and a non-famous user might follow that famous user only due to one aspect. For example, *Lance Armstrong* is famous as a "world-class cyclist" and a "cancer survivor". When a user follows Armstrong, she may be only interested in cycling or she may also be a cancer survivor and interested in charity or she may be interested in both but with different weights. Therefore, they have proposed an LDA-based model named Bi-Labeled LDA to build user implicit interest profiles.

3.2.2 Topic-Centric Models

The intuition behind using the relation between topics of interest to infer implicit interests of users is that a user's interests are semantically or collaboratively related to each other. In other words, if a user is explicitly interested in a topic, one can conclude that the user is probably interested in the topics which are semantically or collaboratively related to her explicit interests. In this section, we review the approaches used in the literature for inferring implicit interests of users by leveraging semantic or collaborative relatedness between topics of interest.

Semantic Relatedness

Inspired by the success of applying a knowledge base in a wide variety of tasks, ranging from recommendation (Wang *et al.*, 2019a,b), dialogue system (Jin *et al.*, 2018), to information extraction (Cao *et al.*, 2018), there is a line of work on implicit user interest mining that first models primitive interests of a user over concepts defined in a knowledge base and then infers the user's implicit interests using the explicit relationships defined in knowledge bases between the concepts. *Primitive interests* here denote the concepts that can be directly spotted from a user's posts (Kapanipathi *et al.*, 2014). Based on the structure of the underlying knowledge base, these studies can be divided into two categories: *hierarchical* and *graph-based*.

Hierarchical Knowledge. In some studies, only the *hierarchical* relations between concepts are considered. For example, given a set of primitive interests of a user which are represented by a set of Wikipedia entities, they utilize Wikipedia category hierarchy to infer high-level interests of the user over Wikipedia categories as the user's hierarchical implicit interests. For instance, in Kapanipathi *et al.* (2014) the authors have first modeled the primitive interests of a user on Twitter by annotating the user's posts with Wikipedia entities and then calculating the degree of the user's interests to each entity by employing a frequency-based scoring mechanism. Then, to extract the hierarchical implicit interests of the user, given the set of scored primitive interests and the Wikipedia category hierarchy, they have first linked the entities in primitive interests of the user to their appropriate categories and considered them as leaf nodes in the Wikipedia category hierarchy. Then, they have calculated the degree of interest of the user to each category by propagating the scores of primitive interests up to the hierarchy as far as the root node using spreading activation theory.

Instead of processing the user's published posts, the authors in Besel *et al.* (2016) have annotated the followees' accounts of a user to Wikipedia entities utilizing MediaWiki Web API[8] to extract her primitive interests. Then to extract the hierarchical implicit interests of the user, given the extracted primitive interests and WiBi (Flati *et al.*, 2014) as a hierarchical knowledge base, they have expanded the user's primitive interests by assigning the entities to categories in WiBi and applying the spreading activation functions proposed in Kapanipathi *et al.* (2014). Similarly, in Faralli *et al.* (2017) the followees of a user have been used to extract the user's primitive interests. However, instead of using all the followees, they have only considered topical users. A topical user is a user who has a corresponding Wikipedia entity. For example, *@britneyspears* is the twitter account of a popular American pop singer described at the link https://en.wikipedia.org/wiki/Britney_Spears. In order to disambiguate user accounts and link them to Wikipedia, the authors have applied Babelfy[9] to the textual content bundled in

[8] https://www.mediawiki.org/wiki/API.
[9] http://babelfy.org/.

each twitter user profile description. Then, to expand a user's interests, they have proposed a methodology to build *Twixonomy*, which is a Wikipedia category taxonomy.

Exploiting hierarchical structure of Wikipedia categories provides a flexible approach through which user interests are able to be represented in multiple levels of granularity (Kapanipathi *et al.*, 2014). However, a major challenge in utilizing Wikipedia category structure as a hierarchy is that it is a cyclic graph instead of a strict hierarchy. This is because categories in Wikipedia are created and edited collaboratively by many different users where any user is free to create or link categories to each other; hence, potentially leading to cyclic references between categories. Using Wikipedia categories without removing cycles would be problematic because cycles make it non-trivial to determine the hierarchical relationships between categories (Boldi and Monti, 2016; Kapanipathi *et al.*, 2014). Therefore, as a preprocess in the user modeling approach, it is required to first transform the Wikipedia category structure into a hierarchy.

To transform the Wikipedia category structure into a hierarchy, for example, the authors in Kapanipathi *et al.* (2014) have removed the Wikipedia admin categories (e.g., Wikipedia, wikiprojects, lists, mediawiki, user, portal, articles and pages) (Ponzetto and Strube, 2007). Then, they have selected *Category:Main_Topic_Classifications*, which subsumes 98% of the categories as the root node of the hierarchy. Next, they have assigned the abstract level of each category based on its shortest path length to the root node. As the last step, all the directed edges from a category of lower hierarchical level (specific) to a higher hierarchical level (conceptually abstract) are removed.

Self-constructed hierarchical structures have been used for mining user interests as well. For example, the authors in Lazzez *et al.* (2018) have constructed user interest profiles based on 24 core Facebook topics of interest (e.g., "Sport and Outdoors", "Food and Drink", and "Shopping and Fashion") based on photos shared by users on Facebook. To this end, computer vision methods such as GoogleNet have been used to classify each photo into 1000 ImageNet concepts. To convert those labels into the 24 core Facebook labels, a hierarchical knowledge base is used. The hierarchical knowledge base is constructed by the

authors where the leaf nodes are the 1000 ImageNet concepts and the 24 core Facebook topics are at the lowest level under the root node (i.e., *User_interest*).

Graph-Based Knowledge. Instead of using *hierarchical* relations between concepts in knowledge bases, some studies have utilized *graph-based* knowledge for inferring the implicit interests of users. Using graph-based knowledge provides more strategies to extend user interests by considering different types of relations between entities and categories. However, different from hierarchical models, the level of granularity of interests is not specified. In the following, we describe some sample studies that utilize a graph-based knowledge for inferring implicit interests of a user.

For example, the authors in Michelson and Macskassy (2010) have first annotated a user's posts on Twitter with Wikipedia articles and then starting from the set of categories on each Wikipedia article, they have traced through the parent categories of each category from Wikipedia category graph which resulted in a set of sub-trees. In the last step, given the forest of sub-trees, they have built the interest profile of a user by discovering the categories that occur frequently and are specific enough to generate useful interest profiles. More specifically, to calculate the degree of interest of the user to each category in the sub-trees, they have considered both the frequency of that category and its level in the sub-trees. Finally, top-k categories are selected as the implicit interest profile of the user.

Similarly, in Orlandi *et al.* (2012) the primitive interests of a user have extended from her published posts by propagating the weights of the user's entities of interest to their associated categories in DBpedia. They have developed two different weighting schemes for calculating the degree of interest of the user to each category. The first approach simply propagates the weights of the user's entities of interest to the categories (i.e., the weight of each category is the sum of all the weights of the entities associated to that category). As the second approach, based on the idea that too broad categories are not descriptive for a user profile, they have proposed a discounting strategy to reduce the

weight of generic categories. Their assumption is that generic categories usually contain many associated entities and have several subcategories.

Instead of using the textual content of the published posts by a user, the authors in Peña *et al.* (2013) have extracted the primitive interests of a user based on the URLs mentioned in the user's posts. For each mentioned URL, if it exists in OpenDNS[10] or DBpedia, they have considered the associated categories to that URL as primitive interests of the user. Then, the user's primitive interests are enriched using similar categories defined by the *sameAs* relationship in DBpedia. For example, the URL https://soundcloud.com/ is tagged into "Music" category by OpenDNS. Searching for this URL in DBpedia results in new categories such as Social_networking_services, Streaming_music_services, Internet_audio_players, Virtual_communities and Music_websites and their *sameAs* categories provided by DBpedia as implicit interests of the user.

While most studies use only the relations between the categories in knowledge bases to expand the primitive interests of a user, the authors in Piao and Breslin (2016c) have investigated other aspects of DBpedia by proposing three propagation strategies: category-based (Figure 3.1(a), class-based (Figure 3.1(b)) and property-based (Figure 3.1(c)). For example, as shown in Figure 4(c), based on the property-based strategy, a user who is interested in Wikipedia entity Steve_Jobs is implicitly interested in Lisa_Brennan-Jobs who is the child of Steve Jobs. Based on their experiments in the context of URL recommendations on Twitter, they have found that expanding user primitive interests by using both categories and related entities, provides the best performance compared to using only categories.

Most recently, researchers have tried to apply knowledge graph information in neural network based embedding methods to improve the performance of user modeling and recommender systems. For example, In Wang *et al.* (2019b) the authors have proposed a method named Knowledge Graph Attention Network (KGAT) to enrich the user-item interactions with high-order item relationships defined on the

[10]OpenDNS cloud websites tagging (http://community.opendns.com/domaintagging/).

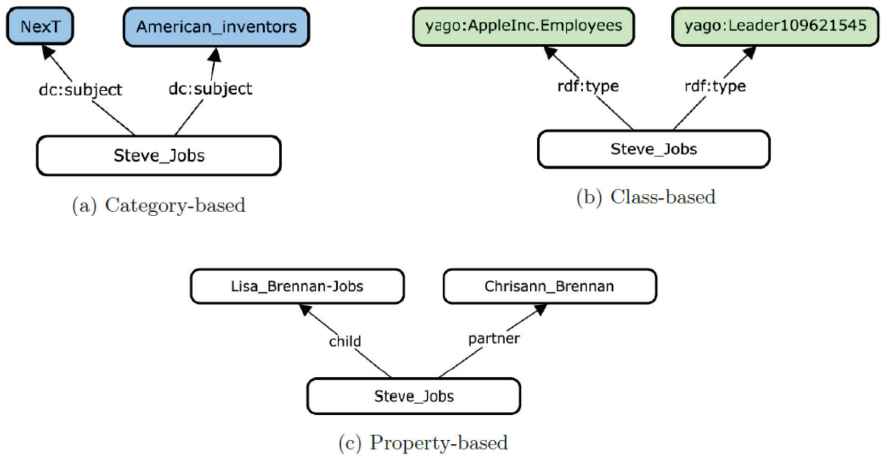

Figure 3.1: Three strategies for user interests using DBpedia (Piao and Breslin, 2016c).

knowledge graph in an end-to-end fashion. Similarly, a deep end-to-end framework is proposed in Wang *et al.* (2019a) which utilizes knowledge graph embeddings to improve recommendations. Most existing studies on knowledge graph enhanced recommendation systems and user modeling methods assume that a knowledge graph is complete and lead to suboptimal performance. This is an unrealistic assumption since it is common that there are some missing relations and concepts in a practical knowledge graph. In Cao *et al.* (2019) the authors have considered the incomplete nature of the knowledge graph when incorporating it into the recommender system. They have proposed a translation-based recommender model which is extended by integrating knowledge graph completion.

Collaborative Relatedness

There is another line of work that instead of using the predefined relations between concepts in knowledge bases, measure the collaborative relatedness between topics based on the overlapping contributions of users over these topics and use them to infer implicit interests of users. For example, in Zarrinkalam *et al.* (2016) the problem of computing the

collaborative relatedness of topics have been regarded as an instance of a model-based collaborative filtering problem (Adomavicius and Tuzhilin, 2005). To this end, they have modeled the explicit interests of users as a user-item rating matrix in which a user-topic entry represents the degree of interest of the user to the topic. By considering the constructed user-item matrix as the ground-truth item recommendation scores, they have learned the relationship between topics in the form of an item similarity matrix by adopting a factored item–item collaborative filtering method (Kabbur *et al.*, 2013) that learns item–item similarities (topic relatedness) as a product of two rank matrices which denote latent factors of items.

As another example, the authors in Trikha *et al.* (2018) have identified users' implicit interests based on topic associations using Frequent Pattern Mining (FPM) without paying attention to the semantics of the topics. Given the collective set of users' explicit interests, they have utilized FPM methods to find the co-occurrence patterns between topics in order to find closely related topics. In other words, they have found all those topics that have frequently co-occurred within the explicit interests of users and expanded the users' explicit interests.

3.2.3 Hybrid Models

While inter-user relationships and inter-topic relationships are both relevant to user interest inference, they have been largely studied in separation in the literature. To take advantage of both indicators, the authors in Piao and Breslin (2017a) have proposed to first leverage biographies of friends of a user to extract the Wikipedia entities of her interests. Based on their idea, a user might be interested in a given topic if she is following who describes herself as a fan or expert of that topic in her biography on Twitter. Then, to expand the implicit interests of the user using inter-topic relationships, they have proposed some propagation methods. The authors have followed a similar approach proposed in Piao and Breslin (2017b) in order to infer implicit interests of passive users by utilizing both user relationships and topic relationships. However, in Piao and Breslin (2017b), they have proposed to

use the information from list memberships of followees instead of their biographies.

Recently, in Zarrinkalam *et al.* (2018) these two factors (i.e., inter-user relationships and inter-topic relationships) are combined into a unified heterogeneous representation model to consider them simultaneously. They have first extracted the users' explicit interests and then proposed to identify their implicit interests by formulating a *graph-based link prediction problem* that operates over a heterogeneous graph by taking into account: (1) user's interests derived from her explicit contributions, (2) relationship between users to incorporate the theory of homophily, and (3) relationship between topics, based on their similar constituent contents and user contributions towards them.

An illustration of the heterogeneous graph is depicted in Figure 3.2. To apply link prediction methods, they have utilized both homogeneous and heterogeneous link prediction methods and shown that heterogeneous approach works better. Based on the result of their experiments, the relatedness between topics is a more accurate clue for inferring implicit interests of users when compared to social relationships, reinforcing the observation that users on Twitter are predominantly interested in semantically related topics. They have also shown that by applying heterogeneous link prediction approach, it is possible to take advantage of both topic relatedness and social relationships simultaneously.

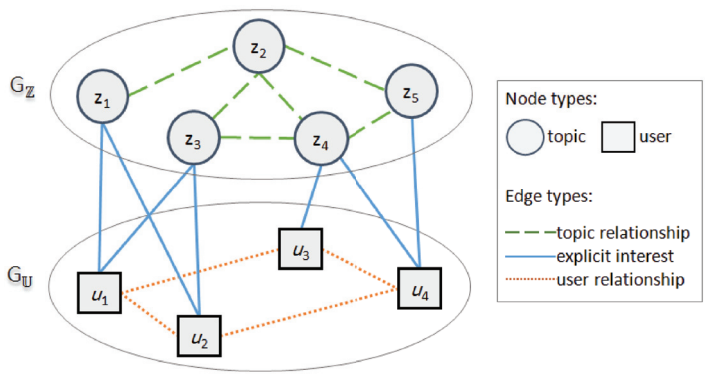

Figure 3.2: An illustration of the underlying graph (Zarrinkalam *et al.*, 2018).

3.2.4 Summary and Discussion

In Subsection 3.2, we reviewed different approaches for implicit interest mining from social media. On one hand, those approaches provide a way of inferring user interests via indirect signals such as following activities or enriching primitive interests using knowledge bases or collaborative knowledge, which is particularly helpful when a user is not actively generating content on social media platforms. On the other hand, leveraging those indirect signals might be noisy and add additional challenges compared to inferring explicit user interests based on their generated content. Table 3.3 summarizes these studies based on the basic user interest modeling aspects introduced in Section 2. Based on this table, the following observations can be made:

- Most studies either use the relationship between users or the relationship between topics for inferring implicit interests of the users, and there are few studies that utilize the potentials of utilizing both types of relationships which results in more accurate implicit interest profiles for users (Zarrinkalam *et al.*, 2018).

- In the studies that focus on detecting implicit interests of the users, social relations mostly follow and retweet relations, and the posts of users are dominantly used as the two internal sources of information. Other types of user content, like images and URLs, are very rarely used.

- The user-centric studies less often use external sources, however, topic-centric studies frequently use external sources, like Wikipedia and DBpedia, for the purpose of automatically identifying the relationships between topics and measuring their relatedness.

- User-centric approaches mainly use the keyword-based techniques for representing implicit interests of the users, while using the concept-based techniques is the norm among the topic-centric and hybrid approaches. This is reasonable since topic-centric and hybrid approaches mainly utilize external knowledge sources for obtaining relationships between topics, and those sources

Table 3.3: Summary of features used for implicit interest mining from social media

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
			User centric				
Pennacchiotti *et al.* (2012)	Posts, Follow Relations		Unigrams				
Wang *et al.* (2013)	Posts, Retweet Relations, Mention Relations, Follow Relations		Unigrams				
Wang *et al.* (2014)	Posts, Retweet Relations, Tags		Group of Keywords (Unigrams)		Sliding Window		
Budak *et al.* (2014)	Posts, Follow Relations, Retweet Relations	DMOZ Taxonomy		DMOZ Categories	Sliding Window		
Bhattacharya *et al.* (2014)	Follow relations, List memberships		Unigrams				
Spasojevic *et al.* (2014)	Posts, Social Relations, Tags	Freebase, Wikipedia		Freebase and Wikipedia Concepts	Sliding Window		Twitter, Facebook, LinkedIn

Continued.

Table 3.3: Continued

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
			User centric				
He *et al.* (2015b)	List memberships, Follow Relations		Tag				
Yin *et al.* (2015)	Tags, Friendship Relations		Group of Keywords (Tags)		Sliding Window		
Grbovic *et al.* (2016b)	Posts, Tags, Follow Relations			Tumblr Taxonomy Concepts	Decay Function		
Gong *et al.* (2020)	Reviews, Posts, Friend Relations		Group of Keywords (Unigrams)				
			Topic centric				
Michelson and Macskassy (2010)	Posts	Wikipedia		Wikipedia Entities and Categories			
Orlandi *et al.* (2012)	Posts	DBpedia, WordNet		DBpedia Entities and Categories	Decay Function	FOAF, WI, WO, OPM, SIOC	Twitter, Facebook

Continued.

Table 3.3: Continued

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
			Topic centric				
Peña *et al.* (2013)	URLs	DBpedia, OpenDNS		DBpedia and OpenDNS Categories		Preference User Profile Ontology (defined in this monograph)	
Kapanipathi *et al.* (2014)	Posts	Wikipedia		Wikipedia Entities and Categories			
Besel *et al.* (2016)	Follow Relations	Wikipedia		DBpedia Entities and Categories		WI	
Piao and Breslin (2016a)	Posts	DBpedia		DBpedia Entities and Categories			Twitter, Google+
Piao and Breslin (2016c)	Posts	DBpedia		DBpedia Entities, Categories and Classes	Sliding Window, Decay Function		
Faralli *et al.* (2017)	Follow relations	Wikipedia, BabelNet		Wikipedia Entities and Categories			

Continued.

Table 3.3: Continued

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
			Topic centric				
Lazzez *et al.* (2018)	Images			Core Facebook Topics	Sliding Window		
Trikha *et al.* (2018)	Posts	Wikipedia		Group of Concepts (Wikipedia Entities)	Sliding Window		
			Hybrid				
Zarrinkalam *et al.* (2016)	Posts, Follow Relations	Wikipedia		Group of Concepts (Wikipedia Entities)	Sliding Window		
Piao and Breslin (2017a)	Follow Relations, Biographies	Wikipedia		Wikipedia Entities and Categories			
Piao and Breslin (2017b)	Posts, Follow Relations	Wikipedia		Wikipedia Entities and Categories			
Zarrinkalam *et al.* (2018)	Posts, Retweet Relations	Wikipedia		Group of Concepts (Wikipedia Entities)	Sliding Window		

organize knowledge in the form of concepts and their semantic, e.g., hierarchical relationships.

- Similar to studies on explicit user interest detection, the temporal aspects of the users' interests have been generally overlooked in the studies that deal with implicit user interest detection. Furthermore, among those studies that take temporal aspects into consideration, the sliding window approach is more commonly used, meaning that the age of users' implicit interests over different time intervals has been ignored.

- The ontologies have been used quite rarely for the purpose of inferring implicit interests, as they have been used only by a few topic-centric studies.

- The potential benefits of using cross-system techniques for detecting implicit interests of the users is largely unexplored in the studies, as only a few studies use information from more than a single social media.

3.3 Future User Interest Prediction

Despite the importance of user future interest prediction on social media, existing studies mainly focus on identifying users' current interests and little work has been done on the prediction of users' potential interests in future. The accurate identification of users' future interests on social media allows one to perform future planning by studying how users will react if certain topics emerge in the future. It leads to improved e-commerce business benefits, targeted advertising and the efficient delivery of services (Bao *et al.*, 2013). For example, it is shown that the box-office revenues of movies can be successfully forecasted in advance of their release by analyzing users' interests in social media (Asur and Huberman, 2010). Therefore, if the forecasted box-office revenues are below the expectations, decision makers can provide film promotion in time for coming up to their expectations.

Dynamicity of users' interests has a key role in future interest prediction and incorporated into future user interest modeling approaches.

Modeling current interests of a user without considering the fact that user interests change over time is not a good estimator to predict her future interests. On social media, not only the user's interests change over time, but also the topics are themselves dynamic in reaction to real-world events. In terms of considering the dynamicity of topics over time, different approaches for future user interest prediction can be divided into two categories: (1) leveraging a fixed topic set, and (2) leveraging a dynamic topic set. In the following, we review the current studies in each category.

3.3.1 Leveraging a Fixed Topic Set

Most studies in future interest prediction assumed that the set of topics stays the same over time and they have predicted the user's degree of interests in future with regard to these topics. For instance, the authors in Gasparetti (2017) have introduced a temporal model that seeks to predict future ratings of a user. Their proposed model, which includes two main steps, quantifies the dynamic interaction between user interest and social trust. In the first step, to model user preferences in each single time point, they have extended Bi-LDA (Porteous *et al.*, 2008) to incorporate social relationships named Bi-LDA^{social}. Then, they have modeled the dynamic interaction of users' interests along with the receptiveness among friends over time by proposing a probabilistic generative model, called Receptiveness over Time Model (RTM).

A temporal and social probabilistic matrix factorization model is proposed in Bao *et al.* (2013) to predict the users' degree of interest in future over a predefined set of trending topics in Sina Weibo. To predict the future interests of a user, their approach is based on the intuition that a user's interests change over time and are affected by the opinions of her friends. To consider the dynamicity of a user's interests, they have divided the historical activities of a user into different time intervals. In each time interval, if the user publishes posts related to a trending topic, it means that she is interested in that topic. After segmenting the historical data of users into different time intervals and modeling the users' interests in each time interval as a user-topic matrix, they have used SocialMF approach, a matrix factorization technique

that also uses the relation between users to learn latent features of users and topics (Jamali and Ester, 2010). Given the latent features of users and topics in each time interval, they have used exponential decay function to estimate the mean matrix of user-latent feature matrix and the mean matrix of topic-latent feature matrix and finally predict the future interests of users.

Similarly, the authors in Arabzadeh *et al.* (2018) have proposed a future interest prediction approach which considers the fact that users' interests change over time by dividing the historical data of users into different time intervals and modeling the user interests in each time interval separately. Their approach is based on the idea that users' future behaviors are influenced by the opinions of other users. Thus, they have utilized Granger causality (Granger, 1969) to find the influencers of a user in a given topic. Finally, to predict the degree of interest of each user to each topic, they have considered the historical interests of the user herself and her top-k influencers in that topic and then apply a vector autoregression.

Despite the above studies which are unsupervised, a supervised method for predicting the users' interests on social media is proposed in Kang *et al.* (2019). They have first applied a word embedding technique to map the words in textual content of social media into vectors and then proposed a deep-neural-network-based approach for predicting user interests over a set of predefined news categories.

3.3.2 Leveraging a Dynamic Topic Set

The approaches mentioned in the previous section assume that the set of topics is fixed over time. This is an unrealistic assumption in the context of social media since the topics rapidly change in reaction to real-world events (Abel *et al.*, 2011a; Huang *et al.*, 2017a). Further, these approaches cannot predict a user's interests with regard to new topics since these topics have never received any feedback from users in the past. Therefore, they suffer from the well-known cold-item problem.

Figure 3.3 illustrates the dynamic occurrence frequency of the main constituent concepts of two topics, i.e., "Jay-Z and Kanye West's performance" and "Teena Marie's death". The occurrence frequencies of

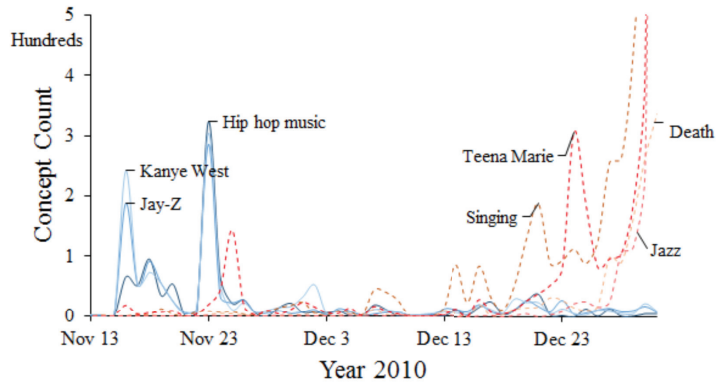

Figure 3.3: Occurrence frequencies of related entities to two real world topics, "Jay-Z and Kanye West's performance" and "Teena Marie's death" (Zarrinkalam *et al.*, 2019b).

these concepts show that "Jay-Z and Kanye West's performance" topic reaches its peaks in mid-November and then decreases rather slowly over the next few weeks, while the latter becomes trendy in late December since spreading the news of Teena Marie's death.

Based on the idea that not only the user's interests change over time, but also the topics are themselves dynamic in social media, the authors in Zarrinkalam *et al.* (2017, 2019b) are among the first to propose a content-based recommender system that utilizes a Wikipedia category hierarchy to serve as a generalizable topic space to predict users' interests over an unobserved set of topics in the future. Their intuition is that although the topics of interests change over time, they are semantically or conceptually similar to a set of core interests. Their approach contains three main steps: (1) they have divided the historical posts of a user into different time intervals and then in each interval, they have applied the LDA method to extract a separate set of topics and the degree of interests of each user to each topic, (2) given the topic profiles of a user in different consecutive time intervals, they have used Wikipedia category hierarchy to model high-level interests of the user by first mapping each topic to its related Wikipedia categories and then applying a spreading activation function to build the hierarchical

category profile of the user, and (3) the future interests of the user are predicted using the extracted category profile.

3.3.3 Summary and Discussion

In Subsection 3.3, we reviewed different approaches proposed for predicting the future interests of users from social media. On one hand, those approaches provide a better understanding of a user's future interests which can be beneficial in many applications such as targeted advertising. On the other hand, similar to explicit user interest mining approaches, the approaches described in this section require users to be actively generating content on social media platforms, which might not be applicable for passive users. Table 3.4 summarizes these studies based on the basic user interest modeling features (introduced in Section 2). Based on this table, the following observations can be made:

- In terms of considering the dynamicity of topics over time, these studies are divided into two categories. In the first category which includes the majority of works, a *fixed set of topics* is considered as the set of topics in the network, and the degree of interest of the users in the future is predicted over the same set of topics observed in the past. However, recently, an approach is presented in Zarrinkalam *et al.* (2019b) for predicting future interests of users over topics unobserved in the past. For this purpose, user interests are modeled over a *dynamic set of topics*.

- Similar to explicit and implicit user interest modeling, users' posts are the main internal source used for predicting future user interests. However, since the future interest prediction from social media is still premature, external sources have been rarely used in the studies.

- Most studies use concept-based techniques for representing future interests of the users, and a few ones employ keyword-based techniques. However, unlike the studies on implicit and explicit user interest detection, the use of Wikipedia and DBpedia is not dominant for detection of future interests of the users.

Table 3.4: Summary of features used for future interest prediction from social media

References	Information Sources		User Interest Representation Units		Temporal Analysis	Ontological Representation	Cross-System
	Internal	External	Keyword-Based	Concept-Based			
				Fixed Topic Set			
Bao *et al.* (2013)	Posts, Follow Relations			Trending Topics in Weibo	Decay Function		
Arabzadeh *et al.* (2018)	Posts		Group of keywords		Sliding Window		
Kang *et al.* (2019)	Posts	News media		News Categories	Sliding Window		
Zheng *et al.* (2019)	Posts, Follow Relations			Predefined Categories	Sliding Window		
				Dynamic Topic Set			
Zarrinkalam *et al.* (2019b)	Posts	Wikipedia		Wikipedia Categories, Group of Concepts (Wikipedia Entities)	Sliding Window		

- Since the goal of the studies in this category is detection of the future interests of the user, all the studies take the temporal aspect of the users' interests into consideration. However, most works are limited to sliding-window techniques and do not consider the interest decaying phenomenon.

- Analysis of the existing studies demonstrates that the ontology based techniques and cross-system approaches are not yet used in the domain of future interest prediction.

- Comparing the Tables 3.1, 3.2 with the Tables 3.3 and 3.4, one can see that the majority of the existing studies in user interest modeling from social media extract the current interests of a user (either explicit or implicit). Little research has been conducted on predicting future interests of users. In addition, considering the recency of the studies on predicting future interests, it can be concluded that this line of research will become more trendy in the near future.

4

Evaluation of User Interest Models

In this section, we first explain different methodologies to evaluate the inferred user interest models and then introduce existing benchmark datasets in the literature for user interest modeling from social media.

4.1 Evaluation Methodologies

In this session, we elaborate on two main approaches used in the literature to evaluate the inferred user interest models, namely *intrinsic* vs. *extrinsic*. The goal of intrinsic approach is to directly evaluate the quality of the constructed user interest profiles by *user study*, a *ground truth* or doing *qualitative analysis* (Chen *et al.*, 2010; Narducci *et al.*, 2013). The extrinsic approach, which has been widely adopted by many researchers, is to evaluate the performance of user interest modeling methods by looking at its impact on the effectiveness of other applications such as news recommendation or tweet recommendation, among others (Abel *et al.*, 2011b; Wang *et al.*, 2014). In the following, we will review how each of these evaluation approaches has been used in the literature.

4.1.1 Intrinsic Evaluation

A common way of evaluating the quality of the inferred user interests is by *user study*, i.e., collecting explicit feedback from users about the inferred interests. The most direct and accurate method for evaluating the inferred interest of a given user via user study is to ask herself.

For example, in order to do user study, the authors in Budak *et al.* (2014) have first randomly selected 500 users among the Twitter users whose profiles include their email addresses. Then, after inferring the top-10 interests of each user, they emailed the user and asked her to evaluate each inferred interest by selecting between Yes or No. Similarly, in Kapanipathi *et al.* (2014), the authors have inferred the interests of 37 volunteers who agreed to participate in user study and evaluate their own interests. They have asked each user to indicate her interest or lack of to each inferred interest by selecting Yes, No or Maybe. The reason for considering Maybe option is that they have modeled user interests over categories, and some identified interests are too abstract and hard for the user to select Yes or No.

Since a small number of users are willing to participate in a user study to evaluate their own inferred interests (e.g., in Budak *et al.*, 2014, only 30 out of 500 selected users accepted to participate in user study), large-scale evaluation is not possible. Therefore, to do a large-scale user study, in some studies, the inferred interests for a given user are evaluated by other users. For example, in Budak *et al.* (2014) a crowdsourcing platform (i.e., Amazon Mechanical Turk platform[1]) is used for evaluation. Although it provides a large-scale evaluation, it is not a very accurate method of evaluation. Because identifying the interests of a user based on her published posts for other users is a hard task and consequently the quality of evaluation results would be questionable at best.

Another intrinsic evaluation methodology is using a *ground truth* to evaluate the inferred interests which requires having a labeled dataset. In some studies, to build a labeled dataset, the authors have asked/hired some human annotators to manually identify the interests of a set of test users based on their activities on social media (Kang *et al.*, 2019;

[1] https://www.mturk.com/.

Kang and Lee, 2017). For example, in Kang and Lee (2017), a labeled dataset for 20 randomly selected users from Facebook is built by asking two human annotators to classify each users' post into, at most, two categories from a predefined set of news categories. The labeled interests of the 20 test users were generated from fully agreed annotations by two annotators.

Instead of totally manually building a ground truth to evaluate the inferred interests, in He *et al.* (2015b), the authors have first used the Stanford POS-Tagger Toutanova *et al.* (2003) to automatically find out all the Twitter users whose biographies contain some predefined patterns (users usually use "play X", "X fan", "interested in X", "love X" or some similar phrases to describe their interests in their biography, where X stands for a topic of interest) and then manually tagged 120 randomly selected users according to their Twitter homepage, biographies and Lists. The reason for manual tagging is that the users can freely express themselves and hence they are ambiguous.

Since human annotators are involved in the process of creating a labeled dataset, it is not possible to prepare a large dataset to do a large-scale evaluation. To solve this issue, some studies (Arabzadeh *et al.*, 2018; Zarrinkalam *et al.*, 2016) have tried to automatically extract explicit interests of users and then consider them as a ground truth to evaluate the inferred users' implicit or future interests. For example, the authors in Zarrinkalam *et al.* (2018) have aggregated the posts published by each Twitter user in a given time interval as a single document and then applied LDA to simply extract the explicit interests of users. Then, based on 10-fold cross validation protocol, they have evaluated the inferred implicit interests of each user extracted by their proposed approach.

In addition to the two aforementioned evaluation approaches, i.e., user study and ground truth, there are some studies that do *qualitative analysis* as an intrinsic approach to evaluate the inferred interests (Besel *et al.*, 2016; Faralli *et al.*, 2017). As an example, in Faralli *et al.* (2017), the authors have performed qualitative analysis to compare their induced user interests against the results of *Who Like What* (Bhattacharya *et al.*, 2014) service and the approach proposed in Kapanipathi *et al.* (2014). They have shown that even though interests are extracted from different

sources (topical users or Twitter Lists), there are several common concepts between the interests inferred by different approaches.

Another example of qualitative analysis is presented in Ottoni *et al.* (2014) where the authors have presented a methodology to study user behavior across multiple social media platforms. The authors collected user activities of about 30,000 users having an account on both Twitter and Pinterest. The authors have performed some quantitative analysis to assess the existence of cross-user behavior patterns.

4.1.2 Extrinsic Evaluation

While the intrinsic approach directly evaluates the quality of the inferred user interests, in an extrinsic evaluation approach, the interests are considered as input in another application and they are evaluated indirectly by evaluating the quality of that application. In other words, the goal of extrinsic evaluation is to investigate if the inferred interests help to solve or improve another problem.

Since in the extrinsic evaluation approach, the human is not usually involved in the evaluation process, large-scale evaluation is possible. However, the comparison of the user modeling strategies is in the context of a *specific* application. For example, if a user interest modeling strategy improves a news recommendation application more than another user interest modeling strategy, it doesn't mean that it is better in general. It is highly probable that we don't get the same result in the context of another application such as friend recommendation. It is because of the fact that different user interest profiles have different levels of performance on different applications (Piao and Breslin, 2018a).

In different studies, different applications such as news recommendation (Abel *et al.*, 2011b), URL recommendation (Piao and Breslin, 2017a), friend recommendation (Kang *et al.*, 2019) and tweet recommendation (Wang *et al.*, 2014) are utilized for evaluating the inferred interests. To apply each of these applications, it is required to answer two main questions: (1) How to obtain the known output of the application? (i.e., what is the ground truth to evaluate the output of the application?). (2) How to incorporate the inferred interests into the application? (i.e., what is the recommendation algorithm?). Given

the ground truth and the recommendation algorithm, we can use the constructed interest profile using a user modeling strategy as an input in the recommender system and then compare the recommendations with the ones in the ground truth to evaluate the quality of the recommendations, and consequently determine how successfully the interests of a user have been identified.

Link recommendation application is frequently used in the literature to evaluate the inferred interests of a user via extrinsic approach (Piao and Breslin, 2016a,c, 2017b). For example, in Piao and Breslin (2017b), the authors have built a ground truth for link recommendation based on the idea that if a user has explicitly mentioned a URL/link in her posts, it means that the user is interested in the web page associated with that link. For the recommendation algorithm, since the main goal is to evaluate the interests, not proposing a new algorithm for link recommendations, they have used a simple algorithm by representing each web page associated with a URL in the candidate set as a weighted vector similar to the user interest profile. Then, for a given user, they have calculated cosine similarity between the vectors of the user's profile and each candidate web page to recommend the most related pages to the user.

In order to evaluate the inferred user interests in the context of a news recommendation application, similar to link prediction application, in Abel *et al.* (2011b), the authors have built a ground truth based on the URLs which are explicitly mentioned in the published posts of users. However, instead of considering all the mentioned URLs, they have only used the URLs that refer to a news article from mainstream news providers such as BBC and CNN. The recommendation algorithm is the same as used for the link recommendation application introduced in Piao and Breslin (2017b).

As another example, the authors in Kang *et al.* (2019) have evaluated the inferred interests in the context of friend recommendation application. To build the ground truth, they have considered the followee of a user on Twitter as her friends. Then, given the constructed interest profile of users as a weighted vector over the topics, by applying cosine similarity, the most similar users to a user are identified and recommended to the user as her friends.

Table 4.1: Summary with basic statistics of single domain benchmarks

Benchmarks	Domain	# users	# items
Yahoo! Music Dataset	Music	1,000 K	625 K
Nantonac	Foods	5 K	100
LARA	Travels	1 K	1.8 K
MovieLens	Movies	280 K	58 K
Million Song	Music	n/a	1,000 K
Netflix Prize	Series, movies, documentaries	480 K	18 K
#nowplaying-RS	Music	139 K	346 K

4.2 Benchmark Datasets

The investigation and diffusion of user interest modeling are both limited by the availability of real user data. Real user data, due to privacy issues, are accessible to a restricted number of researchers. Despite this important limitation, some efforts have been made to create benchmark datasets, by means of dedicated web sites such as Kaggle,[2] where researchers are enabled to publicly release their datasets.

The majority of existing benchmarks (see Table 4.1) are in the general domain of recommender systems and focused only on a single specific area (Çano and Morisio, 2015) such as: music, food and travel, among others. For example, Movielens[3] (Harper and Konstan, 2016) for movies; the Million song dataset[4] (Ly *et al.*, 2018) for music and the Netflix Prize dataset[5] for the on-demand contents of the popular platform (series, movies and documentaries) are large-scale popular benchmarks in the general domain of recommender systems.

Despite the importance of user interest modeling from social media, there is a lack of benchmark dataset to evaluate the inferred user interests. A single domain benchmark database[6] is released in Poddar

[2] https://www.kaggle.com/.
[3] https://grouplens.org/datasets/movielens/latest/.
[4] http://millionsongdataset.com/pages/getting-dataset/.
[5] https://www.kaggle.com/netflix-inc/netflix-prize-data.
[6] The dataset is publicly available under CC BY 4.0 License, https://doi.org/10.5281/zenodo.259453.

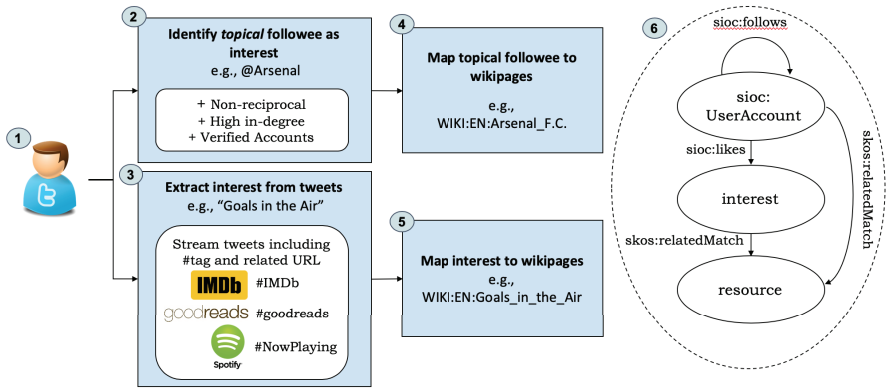

Figure 4.1: WikiMID dataset creation workflow and data model (Tommaso *et al.*, 2018).

et al. (2018) which contains musical interests of users extracted from their Twitter posts. To infer the interests of each user, the authors have only utilized her posts which are automatically posted by the Spotify platform and showing the music that the user is listening to.

Recently, a multi-domain benchmark dataset named WikiMID[7] is built in Tommaso *et al.* (2018) which includes interests of Twitter users from different domains such as music, books and movies. Since WikiMID is the most recent benchmark in the field of user interest modeling, in the following, we explain their approach in more detail. Figure 4.1 shows the workflow used in Tommaso *et al.* (2018) to build the user interest dataset for Twitter users. As it is clearly shown in Figure 4.1, the user interests are extracted based on two independent information sources: user's topical friends (block 2) and users' social posts (block 3).

The authors in Tommaso *et al.* (2018) believe that topical friends of a user are a good source of information to extract the user's interests because users tend to be stable in their relationships and topical friends cover different domains of users' interests such as entertainment, sports and politics. They have distinguished topical friends of a user from her other friends by training a Boolean SVM classifier where the verified

[7]The dataset is publicly available under CC BY NC SA 4.0 License, http://wikimid.tweets.di.uniroma1.it/wikimid/.

Twitter accounts are considered as positive samples. After identifying the topical friends of a user, in the next step (block 4), they have mapped the identified Twitter accounts as user's interests to Wikipedia entities by matching the URL of the Twitter account mentioned in its biography with the URL mentioned in the infobox of the corresponding Wikipedia page.

To reliably extract a user's interests from her posts (block 3), they have not processed all the posts of the user. There are some online platforms that allow users to share their activities and preferences on social media. For example, Spotify allows users to share the music that they are listening to on Twitter by automatically generating a pre-formatted tweet which includes the NowPlaying hashtag followed by the title of music, the name of the artist and finally the corresponding URL. Based on the idea that when a user shares the music, it means she is interested in that music, they have only collected the posts that follow this format for the users. Similar to Spotify, there are other online platforms such as IMDB to share the movies or Goodread that allows sharing the books on Twitter. Therefore, based on some predefined expression, they have only streamed the posts of the users which are shared using these platforms and then extracted the interests of users to different movies, music or books. Given the extracted users' topics of interest, in the next step (block 5) they have mapped each topic of interest to a Wikipedia entity by applying some heuristics and then utilizing ensemble voting to find the most reliable result.

Finally, they have published the extracted user interest dataset using two ontologies, i.e., SIOC and SKOS. The statistics about their published dataset is reported in Table 4.2. Their dataset includes six months data of 2017 in both Italian and English language.

4.3 Evaluation Metrics

In this section we describe the metrics adopted in the literature to evaluate the quality of the inferred user interests using intrinsic or extrinsic evaluation methodologies.

Table 4.2: Structural statistics for the WikiMID dataset

		# interest			
Language	**# user**	**Topical Friends**	**IMDB**	**Goodreads**	**Spotify**
English	444,744	58,780	8,282	20,710	253,311
Italian	25,135	4,580	279	4,690	9,926

4.3.1 Metrics for Intrinsic Evaluation

In intrinsic evaluation methodology, whether the inferred interests are evaluated by the results of a user study or there is a ground truth, based on how the user interest profile is represented, different evaluation metrics are adopted to assess the quality of the inferred interests. In general, given a set of users $\mathbb{U}$ and a predefined set of K topics (e.g., represented by keywords or concepts), denoted by $\mathbb{Z} = \{z_1, z_2, \ldots, z_K\}$, the *interest profile* of a user $u \in \mathbb{U}$, which is the distribution of u's interests over $\mathbb{Z}$, is defined as:

Definition 1 (User Interest Profile). An *interest profile* of a user $u \in \mathbb{U}$, called $P(u)$, is represented by a vector of weights over K topics, i.e., $(f_u(z_1), \ldots, f_u(z_K))$, where $f_u(z_k)$ denotes the degree of u's interest in topic $z_k \in \mathbb{Z}$.

After identifying the interest profile of user u, i.e., $P(u)$, to evaluate the user's inferred interests, some studies (e.g., Abel *et al.*, 2011b; Nishioka and Scherp, 2016; Zarrinkalam *et al.*, 2015) have represented the interests of the user u by a binary representation and then adopted performance metrics for evaluating binary classification problem (summarized in Table 4.3) to evaluate the inferred interests. To represent the inferred interest profile of a user $P(u)$ by a binary representation, two approaches are followed in different studies:

1. Fixing a threshold T on the estimated weight for each topic of interest z_k, i.e., if $f_u(z_k) >= T$, topic z_k is *"of interest"* for user u and if $f_u(z_k) < T$, topic z_k is *"of not interest"* for user u.

Table 4.3: Evaluation metrics in binary classification context

Metric Name	Description
Precision (**P**)	The fraction of correctly inferred interests over all the inferred interests
Recall (**R**)	The fraction of correctly inferred interests over all the user interests in the ground truth
F-Measure (**F**)	The harmonic mean of precision and recall
Accuracy (**Acc**)	The fraction of topics correctly inferred as "of interest" and correctly identified as "of no interest" over the total number of topics in user interest profile $P(u)$
Area Under the Curve (**AUC**)	The trade-off between the true positive rate and false positive rate using different probability thresholds to extract "of interest" and "of not interest" topics.

2. Ranking the topics based on their weights in $P(u)$ and then considering only top-N topics as *"of interest"* and the rest as *"of not interest"*.

Given the binary representation of the constructed user interest profile by a user interest modeling strategy and a set of relevance judgments (i.e., a binary assessment of either "*of interest*" or "*of not interest*" for each user-topic pair), Precision and Recall are two most frequent and basic measures for evaluating the effectiveness of the user interest modeling strategy. In the context of binary classification, the terms True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) compare the results of the user interest modeling strategy under test with the trusted relevance judgments. For example, for a user u, the number of true positives is the number of topics correctly inferred "*of interests*" for u by the user interest modeling strategy. False positives are those topics which are inferred as "*of interests*" for the user u, based on the relevance judgments, but they are not actually among the topics of interests of the user u. Based on

this terminology, Precision (P) and Recall (R) are defined as follows:[8]

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$

It is noted that there are some circumstances in which both Precision and Recall are not applicable. For example, in case of evaluating the constructed user interest profiles by user study, we usually ask the users to evaluate the results of the user interest modeling strategy and therefore we don't have all the actual interests of the user (Kapanipathi *et al.*, 2014). As a result, in such cases, Recall is not applicable and it is only possible to measure the Precision of the results. However, in the cases that both metrics are applicable and important, F-measure is a single measure that trades off Precision versus Recall and is calculated as follows:

$$F = 2\frac{Precision * Recall}{Precision + Recall}$$

Another metric that is used for evaluating a user interest modeling strategy in a binary classification context is Accuracy. Accuracy of each constructed user interest profile, i.e., $P(u)$, measures the fraction of topics correctly inferred as *"of interest"* and correctly identified as *"of no interest"* over the total number of topics in the user interest profile and is calculated as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is an appropriate evaluation metric only when the data is balanced in the user interest modeling problem. However, in most cases the data is unbalanced and a user is interested in a small number of topics in social media. Simply put, for a user most of the data belongs to the *"of not interest"* class. Therefore if a user interest modeling strategy infers that all the topics belong to the *"of not interest"* class for a user, the accuracy of the strategy will be measured as high, but the strategy is not working well.

[8] https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-in-information-retrieval-1.html.

As mentioned before, one approach to represent a user interest profile by a binary representation is fixing a threshold on the estimated topic weights and categorizing the topics into two classes: *"of interest"* and *"of not interest"*. Given the binary representation, it is possible to evaluate the results by metrics such as Precision, Recall, F-Measure and Accuracy. Therefore, the classification threshold value has an impact on the evaluation results by these metrics and needs to be analysed. Despite these metrics, AUC measures the quality of the model's predictions irrespective of what classification threshold is chosen. AUC is the area under the ROC curve (Receiver Operating Characteristic curve) which plots True Positive Rate vs. False Positive Rate at different classification thresholds. True Positive Rate (TPR) is a synonym for Recall and is therefore measured by $\frac{TP}{TP+FN}$ and False Positive Rate (FPR) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

Most studies (e.g., Abel *et al.*, 2011b; Nishioka and Scherp, 2016; Zarrinkalam *et al.*, 2015), have represented the topics of interest, $\mathbb{Z} = \{z_1, z_2, \ldots, z_K\}$, as a ranked list, ordered descendingly based on their weights in $P(u)$, i.e., $(f_u(z_1), \ldots, f_u(z_K))$ and then used the standard information retrieval metrics for evaluating the ranking quality of the inferred interests. These metrics are summarized in Table 4.4.

As mentioned before, Precision and Recall are set-based measures and do not consider the order of results. Therefore, to evaluate the ranked retrieval results, these metrics are extended to $P@N$ and $R@N$ by applying over the top N retrieved interests. Similarly $S@N$ measures the mean probability that a relevant item occurs within the top N of the recommendations. However, these metrics fail to take into account the positions of the relevant interests among the top N. Therefore, to better evaluate the ranking of results, other metrics such as Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG) have become more common for evaluating the user modeling strategies which results in a ranked list of user interests.

For a single user, let the Average Precision (AP) be the average of the precision value obtained in different Recall values. To measure

Table 4.4: Metrics for evaluating the ranking quality

Metric Name	Description
Precision at rank N (**P@N**)	The fraction of correctly inferred interests to all the inferred interests
Recall at rank N (**R@N**)	The mean probability that relevant topics are successfully retrieved within the top-N recommendations
Success at rank N (**S@N**)	The mean probability that relevant topics occurs within the top-N recommendations
Mean Average Precision (**MAP**)	How well the interests are ranked at top-N and how early relevant results appear
Mean Reciprocal Rank (**MRR**)	At which rank the first topic relevant to the user occurs on average
Normalized Discounted Cumulative Gain (**NDCG**)	How well highly relevant topics are appeared earlier in the ranked list of topics

MAP, the value of AP is averaged over all users. Therefore, given $\mathbb{Z} = \{z_1, z_2, \ldots, z_K\}$ as a ranked list to represent the inferred topics if interests of a user, MAP is calculated as follows:

$$MAP = \frac{1}{|\mathbb{U}|} \sum_{i=1}^{|\mathbb{U}|} \frac{1}{K} \sum_{k=1}^{K} P@k * rel(k)$$

where $rel(k)$ is equal to 1 if the topic at rank k is among the actual topics of interest of user u and otherwise it is 0.

MRR and NDCG are also two alternative evaluation metrics which are used in some studies in addition to MAP to evaluate the ranking of inferred interests. Like $P@N$, NDCG is evaluated over top-N results. For a set of users $\mathbb{U}$, let rel_j be the relevance score assessors gave to the interest at position j. Then, NDCG is calculated as follows:

$$NDCG@N = \frac{1}{|\mathbb{U}|} \sum_{i=1}^{|\mathbb{U}|} Z_{Ni} \sum_{j=1}^{N} \frac{2^{rel_j} - 1}{\log_2(j+1)}$$

where Z_{Ni} is a normalization factor calculated to make it so that a perfect ranking's NDCG at N for user u_i is 1.

Similarly, MRR is calculated as follows:

$$MRR = \frac{1}{|\mathbb{U}|} \sum_{i=1}^{|\mathbb{U}|} \frac{1}{Rank_i}$$

where $Rank_i$ refers to the rank position of the first topic in the retrieved ranked list of user interests which is *"of interest"* for user u_i.

Although the aforementioned metrics are adopted in most of the user interest modeling approaches for intrinsic evaluation, given a ground truth of user interests with the real values of user's interests, it is possible to measure Mean Absolute Error (MAE) and Root-Mean-Squared Error (RMSE) to evaluate the inferred interests. For example, in Arabzadeh *et al.* (2018), the authors have evaluated the predicted future interests of users by MAE and RMSE metrics.

4.3.2 Measures for Extrinsic Evaluation

In contrast to intrinsic evaluation, through extrinsic evaluation methodology the performance of a system is estimated in the context of an application, where the output of the application depends on the output of the system.

In Subsection 4.1.2, we discussed how extrinsic evaluation methodology is adopted to evaluate user interest modeling strategies, and showed that the majority of them involves the evaluation in the context of recommender system applications. A recommender system operates starting from user interests, and generates a ranked list of items to recommend as "*of interest*" to a specific user. Therefore, to indirectly evaluate the quality of user interest modeling strategy, given the actual users' items of interests, in most studies, the performance of the recommender system application is estimated using the quality ranking metrics, summarized in Table 4.4 (e.g., Piao and Breslin, 2017b; Trikha *et al.*, 2018; Zarrinkalam *et al.*, 2018).

Perplexity is also adopted in some studies (e.g., Tang and Yang, 2012; Wang *et al.*, 2014; Xu *et al.*, 2011; Zarrinkalam *et al.*, 2018), to evaluate the overall generalizability of modeling unseen/implicit user interests in each comparison method. Perplexity is widely used in language modeling and topic modeling to evaluate the predictive power

of a topic model (Blei *et al.*, 2003; Liu *et al.*, 2015; Mao *et al.*, 2012; Steyvers *et al.*, 2004).

Perplexity is equivalent algebraically to the inverse of the geometric mean of per word likelihood (Blei *et al.*, 2003). A good topic model should be able to generalize to unseen data and a lower perplexity score indicates better generalization performance. Formally, for a test set of documents, denoted by D_{test}, the perplexity is calculated as follows:

$$perplexity(D_{test}) = \exp -\frac{\sum_{d \in D_{test}} \log(p(W_d))}{\sum_{d \in D_{test}} |W_d|}$$

where W_d represents the words in document d and $p(W_d)$ is the probability of test document d. Let $\mathbb{Z}$ be the set of topics as the output of topic model and $p(w, z)$ denotes the probability of word w in topic z, $p(W_d)$ is calculated as:

$$p(W_d) = \sum_{z \in \mathbb{Z}} \prod_{w \in W_d} p(w, z)$$

Only a few of user modeling strategies that use a latent topic model framework to represent user interests as a distribution over latent topics have utilized perplexity to evaluate the generalizability performance of their model. For example, in Xu *et al.* (2012), the authors have proposed a mixture latent topic model framework which has a similar general structure to the author-topic model (Steyvers *et al.*, 2004) by considering the tweets published by a user as a document. However, despite the author-topic model, they have considered three more factors: breaking news, posts from social friends and user's intrinsic interest. To evaluate the generalization performance of their mixture topic model, they have randomly split the posts of each user into 90% training posts to infer the topics of interest of the user and 10% test posts, and computed the perplexity of all test tweets according to perplexity metric.

As other examples, in Wang *et al.* (2014) and Zarrinkalam *et al.* (2018), the authors have utilized perplexity to evaluate the generalization ability of their model. However, in contrast to the approach followed in Xu *et al.* (2011, 2012), instead of considering each post as a document, they have aggregated the posts published by a user in a given time interval as a document to build each document in the test set.

4.4 Summary and Discussion

In this section, we reviewed different evaluation methodologies, existing benchmark datasets and metrics to evaluate the inferred user interests. Two methodologies are mainly used to evaluate the inferred user interest profiles: (1) intrinsic approach, i.e., directly evaluating the quality of the constructed user interest profiles by user study or ground truth, and (2) extrinsic approach, i.e., evaluating the performance of user interest modeling methods by looking at its impact on the effectiveness of other applications such as news recommendation or tweet recommendation, among others. An intrinsic approach could achieve more accurate evaluations with direct feedback from users but is usually limited by the number of users for evaluation. In contrast, an extrinsic approach can provide a larger scale evaluation compared to the intrinsic one via indirect evaluation such as news or tweet recommendation performance. However, it might not be as accurate as the intrinsic evaluation.

Whether intrinsic or extrinsic, since it's hard to collect negative samples for negative testing, most of the existing studies in the literature have only evaluated their proposed user interest modeling strategies by positive test cases and utilized evaluation metrics such as Precision, MAP and NDCG. The main reason is that the majority of social media platforms do not provide a way of collecting negative test cases to build a ground truth, e.g., via a "dislike" button with an exception for YouTube. For example, a common practice for extrinsic evaluation, e.g., via tweet recommendations is using tweets that are retweeted by a user as positive cases to build the ground truth to evaluate the output of the application and indirectly evaluate the performance of the user interest modeling strategy considered as an input in the application. It is not a reasonably accurate assumption to assume the other tweets from her followees' timeline, which are not tweeted by the user during the retweeting timespan, as negative cases.

In Section 3, different studies in the field of user interest modeling are divided into three main categories, i.e., *explicit user interest detection*, *implicit user interest mining* and *future user interest prediction*. Tables 4.5–4.8 summarize the studies introduced in Section 3 in

each category in terms of their evaluation methodologies, datasets and metrics. Based on these tables, the following observations can be made:

- Due to the possibility of doing large-scale evaluation via extrinsic evaluation, it is the most common and dominant evaluation methodology for indirectly evaluating the quality explicit/implicit user interest modeling strategies by the quality of the recommender system application for which the inferred interests are used.

- Although there are some studies that only evaluate their explicit/implicit user interest modeling strategies by qualitative analysis, more recent studies use these analyses as a complementary approach along with another evaluation approach such as user study or extrinsic.

- To evaluate the future interest prediction methods, all the studies use a ground truth. As mentioned in Subsection 3.3, future interest prediction methods work based on the historical interests of users modeled over different time intervals. Therefore, to evaluate their methods, given N time intervals of the historical interests of users, they have considered the first $N-1$ time intervals for training and interest profile of users in the last time interval as a ground truth for testing.

- Most of the studies in all the three types of user interest modeling strategies use Twitter as their source of information, because the information that the users publish on Twitter are more publicly accessible compared to other social media.

- Due to the lack of a comprehensive user interest dataset or benchmark for evaluation, different studies have used different datasets which resulted in many challenges in the comparison. Further, some works do not explicitly describe with enough details all preprocessing steps applied to the adopted datasets. Therefore, although there are many studies in this field they don't compare their approach with each other.

Table 4.5: Summary of evaluation methodologies, datasets and metrics used for explicit interest detection from social media (part 1)

References	Evaluation Methodology		Dataset			Metric
	Intrinsic	**Extrinsic**	**Social Media**	**Collection Size**	**Temporal Scope**	
			Frequency-based			
Wang *et al.* (2009)		Ad recommendation	Flickr	25 Flickr users 5 K photos		MAP
Abel *et al.* (2010)	Qualitative		Twitter, Blogspot, Flickr, Delicious, StumbleUpon	30 users 159 tags		Precision
Abel *et al.* (2011c)	Qualitative		Twitter	1 K users 3.3 M tweets		
Abel *et al.* (2011b)		News recommendation	Twitter	1.6 K users, 2 M tweets	2 months	S@10, MRR
Kapanipathi *et al.* (2011)	Qualitative		Twitter, Facebook, LinkedIn			
Gao *et al.* (2012)	Qualitative	Tweet recommendation	Twitter	73 users, 41 K tweets	Jan. 1st to Jul. 7th, 2011	MRR
Abel *et al.* (2011a)	Qualitative	Link Recommendation	Twitter	20 K users, 30 M tweets	4 months (from Nov. 15th 2010)	S@10, MRR
Yang *et al.* (2012)		Hashtag recommendation	Twitter	Dataset 1: 373 K tweets, 1 K users, 9 K hashtags Dataset 2: 19 M users, 49 M hashtags, 476 M tweets	Dataset 1: Mar 25th 2007 to Dec. 13th 2010 Dataset 2: Jun. 1st 2009 to Dec. 31st 2009	Accuracy

Continued.

Table 4.5: Continued

References	Evaluation Methodology		Dataset			Metric
	Intrinsic	**Extrinsic**	**Social Media**	**Collection Size**	**Temporal Scope**	
			Frequency-based			
Abel *et al.* (2013b)	Qualitative	Tag recommendation, Resource recommendation	Flickr, Delicious, and StumbleUpon, Facebook, Twitter, LinkedIn	321 users		Accuracy, MRR, S@N, P@N
Joshi *et al.* (2015)	Qualitative		Flickr	2 K users, 1 M images		
Cinar *et al.* (2015)	Ground truth		Pinterest	650 users, 1 M pins,18 K boards		Accuracy, R@N
Zarrinkalam *et al.* (2015)		News recommendation	Twitter	135 K users, 3 M tweets	Nov. 1st to Dec. 31st, 2010	MRR, S@N
Elmongui *et al.* (2015)		Tweet recommendation	Twitter	20 K users ,9.1 M follow relations, 10 M tweets		NDCG, MAP
Nishioka and Scherp (2016)	Qualitative	Publication Recommendation	Twitter	123 users, average 1 K tweets per user		MAP, MRR, NDCG

Continued.

Table 4.5: Continued

References	Evaluation Methodology		Dataset			Metric
	Intrinsic	**Extrinsic**	**Social Media**	**Collection Size**	**Temporal Scope**	
			Frequency-based			
Tommaso *et al.* (2018)	Qualitative		Twitter	En dataset: 444 K users, 2.7 M tweets It dataset: 25 K users, 84 K tweets	Apr. to Sep. 1st, 2017	
Wieczorek *et al.* (2018)		News recommendation	Twitter	135 K users 3 M tweets	Nov. 1st and Dec. 31st, 2010	MRR, S@N
Nguyen *et al.* (2016)	Qualitative		Reddit	2 well-known users		
Dinh and Van Pham (2020)	Ground truth		Facebook	256 users		Accuracy, Precision, Recall, F-measure
Pandey and Sang (2015)	Ground truth		Instagram	50 users 2,600 photos		Accuracy

Table 4.6: Summary of evaluation methodologies, datasets and metrics used for explicit interest detection from social media (part 2)

References	Evaluation Methodology		Dataset			Metric
	Intrinsic	**Extrinsic**	**Social Media**	**Collection Size**	**Temporal Scope**	
			Probabilistic			
Weng *et al.* (2010)	Qualitative		Sina Weibo	20 users, 10 K posts		Perplexity
Xu *et al.* (2011)	Qualitative	Unseen-data prediction	Twitter	2.62 M tweets 16 K users	2 months	Perplexity
Ahmed *et al.* (2011)		Ad recommendation	Yahoo	Dataset 1: 13 M users, 100 K vocabulary Dataset 2: 33 M users, 100 K vocabulary	Dataset 1: 56 days Dataset 2: 44 days	ROC
Ottoni *et al.* (2014)	Qualitative		Twitter, Pinterest	23 K users, 1.8 M pins, 4.4 M tweets	Sep. 12th to Dec. 3rd, 2013	
Khater *et al.* (2014)		Tweet recommendation	Twitter	20 K users, 5 M tweets	Apr. to Jun. 2013	Precision, Recall
Yao *et al.* (2017)	Qualitative	Friend recommendation	Flickr	10 K users, 2.4 M images, 20 K words	MRR	
Yin *et al.* (2015)		News, tag recommendation	Delicious, Digg	Delicious: 201 K users 2.8 M tags Digg: 139 K users 3.5 K news		Precision, NDCG, F-measure
Sang *et al.* (2015)		Retweet prediction News recommendation	Twitter	19.2 K users		Accuracy, MAP

Continued.

Table 4.6: Continued

References	Evaluation Methodology		Dataset			Metric
	Intrinsic	**Extrinsic**	**Social Media**	**Collection Size**	**Temporal Scope**	
ML-based						
Zhao *et al.* (2015)		Unseen-data prediction	Google+		May and Jun. 2014	NDCG@N, Recall@N, Average Percentile@N
Cao *et al.* (2017)		Tag score prediction	Twitter	10 K users		Precision@N, MAE
You *et al.* (2016)	Ground truth		Pinterest	748 users 30,213 pinboards 1,586,947 pins		NDCG, Recall, Accuracy
Geng *et al.* (2014)	Ground truth		Pinterest	1.2 K users 1.5 M photos		MAP@N, F-measure
Fiallos and Jimenes (2019)	Ground truth		Twitter	1,573 users 100 tweets per user on average		Precision, Recall
Xu and Lu (2015)		Unseen-data prediction	Tumblr	1,086 topics	Jun. 1st and Jun. 30th, 2014	MAP
Similarity-based						
Kang and Lee (2017)	Ground truth		Facebook	7.8 K posts 20 users	Accuracy	
Liang *et al.* (2018)	Ground truth		Twitter	1.3 K users 3.7 M tweets	For each user, from the beginning of her registration up to May 31st, 2015	P@N, NDCG@N, MRR@N, MAP@N
Seghouani *et al.* (2018)	Ground truth		Twitter	Multilingual dataset: 1.3 K users, 3 M tweets Monolingual dataset: 1.4 K users, 600 K tweets		Precision, Recall, F-measure

Continued.

Table 4.7: Summary of evaluation methodologies, datasets and metrics used for implicit interest mining from social media

References	Evaluation Methodology		Dataset			Metric
	Intrinsic	Extrinsic	Social Media	Collection Size	Temporal Scope	
User centric						
Pennacchiotti *et al.* (2012)		Tweet recommendation	Twitter	182 K tweets, 43 K hashtags	Oct. 30th and Nov. 4th, 2011	P@N, S@N, MRR
Wang *et al.* (2013)	Ground truth		Weibo	5.2 K users, 134 K follow relation, 22.3 K keywords		NDCG
Wang *et al.* (2014)	Ground truth	Tweet recommendation, Finding topical authorities, Unseen data prediction	Twitter	28 K users 19 M tweets 1.2 M follow relation 415 K retweet relation	Aug and Sep. 2009	Perplexity P@N MRR
Budak *et al.* (2014)	User study		Twitter	500 users Each user at least 50 tweets	May 2011 - May 2012	Precision
Bhattacharya *et al.* (2014)	Qualitative, User study		Twitter	10 users		Accuracy
Spasojevic *et al.* (2014)	User study		Twitter, Facebook, Google Plus, LinkedIn	43 users 32.2 K (user, topic) labels		Precision, Recall, F-measure
He *et al.* (2015b)	Qualitative, Ground truth		Twitter	26.4 K users 2.7 M follow relations 159 tags	Jun. 1st, 2009 to Dec. 31st, 2009	Recall, Precision

Continued.

Table 4.7: Continued

References	Evaluation Methodology		Dataset			Metric
	Intrinsic	Extrinsic	Social Media	Collection Size	Temporal Scope	
User centric						
Yin *et al.* (2015)	Qualitative	Item recommendation	Digg, MovieLens, Douban Movie and Delicious	Digg: 139 K users, 3 M votes MovieLens: 33 K users, 87 K movies Douban Movie: 71 K users, 10 K movies Delicious: 201 K users, 2 M tags	Digg: 2009–2010 MovieLens: 1998–2009 Douban Movie: 2005–2010 Delicious: 2008–2009	Precision, NDCG, F-measure
Grbovic *et al.* (2016b)	Qualitative	Ad engagement	Tumblr	96.9 M users 26.1B activities		A/B testing
Gong *et al.* (2020)		Link prediction	Yelp	10.8 K users 187 K reviews		Perplexity, NDCG, MAP
Topic centric						
Michelson and Macskassy (2010)	User study		Twitter	4 users 1.3 K tweets		MAP
Orlandi *et al.* (2012)	User Study		Twitter, Facebook	21 users		P@10, MRR,
Peña *et al.* (2013)	Qualitative		Twitter	18 K tweets 8 K users		
Kapanipathi *et al.* (2014)	User Study		Twitter	37 users 31.9 K tweets		MAP, MRR
Besel *et al.* (2016)	Qualitative, User study		Twitter	3 K users 350 K followees		MAP, MRR
Piao and Breslin (2016a)		Link recommendation	Twitter and Google+	429 users		MRR, Recall
Piao and Breslin (2016c)		Link recommendation	Twitter	480 users 348 K tweets		MRR, S@N, R@N and P@N

Continued.

Table 4.7: Continued

References	Evaluation Methodology		Dataset			Metric
	Intrinsic	Extrinsic	Social Media	Collection Size	Temporal Scope	
			Topic centric			
Faralli *et al.* (2017)	Qualitative		Twitter	Dataset 1: 41 M users, 40 M followees Dataset 2: 101 K users, 16.3 M follow relation	Dataset 1: Jun. to Aug. 2009 Dataset 2: Mar. 2014 to Jun. 2014	
Lazzez *et al.* (2018)	Qualitative, Ground truth		Facebook	240 users 24 K images		Accuracy
Trikha *et al.* (2018)		Tweet recommendation	Twitter	3 M tweets 135 K users	Nov. 1st to Dec. 31st, 2010	MAP
			Hybrid			
Zarrinkalam *et al.* (2016)	Ground truth		Twitter	3 M tweets 135 K users	Nov. 1st to Dec. 31st, 2010	AUPR, AUROC
Piao and Breslin (2017a)		Link recommendation	Twitter	50 users 72 K followees with bio		MRR, S@N, R@N and P@N
Piao and Breslin (2017b)	Qualitative	Link recommendation	Twitter	Dataset 1: 480 users 170 avg number of followees Dataset 2: 173 avg number of list memberships of followees		MRR, S@N, R@N, P@N
Zarrinkalam *et al.* (2018)	Ground truth	Tweet recommendation, Unseen data prediction	Twitter	130 K users 3 M tweets	Nov. 1st to Dec. 31st, 2010	AUROC, Perplexity, MAP, NDCG@N

Table 4.8: Summary of evaluation methodologies, datasets and metrics used for future interest prediction from social media

References	Evaluation Methodology		Dataset			Metric
	Intrinsic	Extrinsic	Social Media	Collection Size	Temporal Scope	
Fixed Topic Set						
Bao *et al.* (2013)	Ground truth		Weibo	1 K users 2.7 K topics	Oct. 29th, 2011 to Nov. 13th, 2011	P@N
Arabzadeh *et al.* (2018)	Ground truth		Twitter	2 M tweets 135 K users	Nov. 1st to Dec. 31st, 2010	MAE, RMSE, NDCG, MAP, P@N
Kang *et al.* (2019)	Ground truth		Twitter, Facebook	Facebook: 7.8 K posts, 20 users Twitter: 43 K posts, 70 users		Accuracy
Zheng *et al.* (2019)	Ground truth		Weibo	30 K users 17 M posts	Jan. 20th, 2017 to Apr. 1st, 2017	Precision, Recall, F-Measure
Dynamic Topic Set						
Zarrinkalam *et al.* (2019b)	Ground truth		Twitter	2.5 M tweets 2.4 K users	Nov. 1st to Dec. 31st, 2010	NDCG, MAP

- Most studies do not provide a systematic discussion of the error committed by the proposed modeling strategy.
- In most studies, the experiments are done only on a small dataset, for example, only 50 users or based on only three million tweets. Therefore, evaluating the scalability of use interest modeling strategies is overlooked in most of the studies.
- Most of the studies use only one dataset in a given time interval. Given that the characteristics of social media data such as user population and user posting behavior may have changed over time (Liu *et al.*, 2014), the influence of Twitter evolution on the performance of their proposed model and the validity of their findings on different times are not investigated.
- In terms of evaluation metrics, whether intrinsic evaluation or extrinsic evaluation, since the ranking of results is important in most studies, the metrics that estimate the quality of rankings are mostly dominant metrics in the field of user interest modeling.

5

Applications of User Interest Models

In this section, we introduce different types of applications that have been taking advantage of user interest modeling from social media platforms to improve their services.

5.1 Applications on Social Media Platforms

User-generated content is the lifeblood of social media platforms. With the availability of user-generated content and the ability of user interest modeling approaches to infer user interest profiles, social media platforms themselves can provide many useful applications to their users. In the following, we provide an overview of some of the applications on social media platforms that can take advantage of user interest modeling.

5.1.1 User-Aware Recommendations

One of the most important applications of social media platforms is recommending personalized content or friends to the users by awaring their interests in order to increase their engagement with other users and keep them generating content on these platforms.

Personalized Content Ranking and Recommendations. A typical example on Twitter is recommending a (ranked) list of tweets or topical lists that a user might be interested in from a large amount of content on Twitter, in order to increase the possibility of liking or re-sharing the content. For example, keyword-based user interest profiles have been used for the application of tweet recommendation. In Feng and Wang (2013), the authors have used TF-IDF weighted words based on all and recent tweets of a user to represent long- and short-term user interest profiles, and those interests are used as features in their proposed Feature-aware Factorization Model to provide personalized ranking of tweets for users.

The authors in Sang *et al.* (2015) have proposed a probabilistic topic model for temporal user modeling on microblogs to infer both long- and short-term user interests. User interests are represented by topics (group of keywords) and long-term interests here denote the general topics in a global timeline while short-term topics denote temporal topics discovered at each corresponding time interval. The inferred user interest profiles are used for recommending tweets that would be retweeted by a user and providing personalized news recommendations. Similarly, to recommend tweets that a user would retweet in the future, in Hong *et al.* (2013), the authors have used Factorization Machines where users' topics of interest are extracted by learning a topic matrix, which is similar to traditional topic models such as Latent Dirichlet Allocation (LDA).

Concept-based interest profiles have been used for recommending content as well. For example, the authors in Karatay and Karagoz (2015) have used concept-based user profiles where a concept is a named entity that can be found in Wikipedia for recommending tweets to users. In Lu *et al.* (2012) an approach is proposed for mining user interests from Twitter where the explicit user interests are represented by weighted Wikipedia concepts which are extracted from a user's tweets. The concept-based profiles can be further enriched by her implicit interests by exploring the Wikipedia concept graph based on the rich inter-link information between concepts. Afterwards, tweets in a user's timeline on Twitter are re-ranked based on the similarity between each tweet and the user's interest profile. As another content recommendation

application, to provide personalized recommendation of Twitter lists, in Rakesh *et al.* (2014) a temporal topic-preference matrix is learnt to capture the inclination of users towards a set of topics at different time frames using discrete dynamic topic model (dDTM) (Blei and Lafferty, 2006).

Social media content can consist of different languages or different types of content such as text or images. Therefore, multi-lingual and multi-modal content recommendations also have been considered in previous studies. For example, Seghouani *et al.* (2018) have proposed a Wikipedia-based approach to cope with multilingual texts on Twitter for inferring user interests. To this end, TagMe is used to extract Wikipedia entities in four different languages from tweets. Then, their corresponding categories are inferred and being used as bag-of-concepts to represent each user interest.

Recently, Pinterest developed "Pinterest Taxonomy" which is a taxonomy to represent user interests. This taxonomy supports different languages with English version as a ground truth to help both content recommendation and ads targeting (Gonçalves *et al.*, 2019). User interests can be inferred based on their pins and Pin2Interest[1] which is a machine learning system which maps pins to interests/nodes in the "Pinterest Taxonomy" by exploring multi-modal information associated with each pin such as images, annotations, etc. Multi-modal approaches have shown their effectiveness in predicting user interests. For example, in Cinar *et al.* (2015), the authors have shown that their multimodal approach, which considers both text and images, outperforms predicting user interests on Pinterest compared to unimodal approaches.

Personalized Friend Recommendations. Another important characteristic of social media platforms is the ability to enable each user to establish connections with other users for sharing interests, activities, backgrounds and real-life relationships. The significance of resulting user social ego-networks plays an important role for fulfilling users' needs and expectations on those social media platforms such as: social

[1] https://medium.com/pinterest-engineering/pin2interest-a-scalable-system-for-for-content-classification-41a586675ee7.

interaction, seeking information, entertaining and sharing information (Anita, 2013). Therefore, recommending right connections to users plays a crucial role for keeping users staying and using those platforms. To this end, many applications have been proposed for providing personalized friend recommendations on different social media platforms such as Twitter, Flickr and Facebook (Huang *et al.*, 2017b; Naruchitparames *et al.*, 2011; Weng *et al.*, 2010).

For example, in Hannon *et al.* (2010, 2011b), the authors have built keyword-based user interest profiles based on the tweets of each user or her followers and followees on Twitter. All those tweets (relevant to a user) are considered as a document and each term in the document is given its importance using the TF-IDF score. Afterwards, a user can enter search queries to find relevant users (potential friends) based on the similarity between the queries and the user documents in the corpus. The authors in Zheng *et al.* (2015) have proposed a method for mining user interest profiles from Sina Weibo. The authors claimed that using bag-of-words approaches are not sufficient for discovering users' interests. In contrast to bag-of-words, in order to incorporate the temporal dynamics of user interests, the proposed method leverages a set of microblogs published by a user on a specific time interval, and then applies LDA to learn the topics that the user is interested in during that time period.

As another example for friend recommendation application, the authors in Faralli *et al.* (2015) have used a taxonomy of Wikipedia categories called "Twixonomy" to build hierarchical interest profiles where user interests are represented as Wikipedia categories. Those inferred concept-based interest profiles are then applied for recommending user Twitter accounts that a user might be interested in. Similar to the trend of using multi-modal approaches for content recommendations, in Yao *et al.* (2017), it is shown that multimodal approaches are also helpful for improving the performance of friend recommendations by exploring images and tags together on Flickr.

5.1.2 Advertisement (Ad) Recommendations

Advertising is a critical source where social media platforms derive their revenue (Rohn, 2015). Facebook alone, for instance, generated $17.4 billion ad revenue in the third quarter of 2019.[2] Hence, targeted advertising, which is a type of online advertising that focuses on different aspects of a user such as interests or gender, has become one of the largest and most lucrative advertising channels (Grbovic *et al.*, 2016a).

For example, the "Pinterest Taxonomy" (aforementioned in Subsection 5.1.1) also has been used in Pinterest for ad targeting and recommendations. A machine learning (ML) system called User2Interest is developed to map users to their interests in the taxonomy, which uses user engaged pins and corresponding interest labels of those pins output from Pin2Interest as major input. This can provide statistics such as the number of users per taxonomy node which can be used for informing advertisers of shifts in overall interest.[3]

Tumblr also provides similar interest/topic taxonomy for advertisers to provide sponsored post advertising (Grbovic *et al.*, 2015, 2016b). User interests are mapped to topics in a two-level General Interest Taxonomy (GIT), which is used by the Yahoo Gemini[4] advertising. The GIT is carefully derived based on Interactive Advertising Bureau (IAB) taxonomy recommendations,[5] in order to meet advertiser needs and protect Yahoo's interests (Grbovic *et al.*, 2016b). To this end, the authors proposed a semi-supervised skip-gram modeling approach (Mikolov *et al.*, 2013) to learn tags and categories (topics in GIT) associated with Tumblr posts. Afterwards, user interest profiles are built based on their activities (e.g., generated posts, following other users) and time for applying interest decay.

Similar approaches that map users to interest labels – so that advertisers can target relevant users for advertising – can be found in other social media platforms as well. The "Interest Targeting" service

[2] https://adage.com/article/digital/facebook-reports-increases-ad-revenue-and-users-third-quarter/2211401.

[3] https://venturebeat.com/2020/01/10/pinterest-details-the-ai-and-taxonomy-systems-underpinning-its-trends-feature/.

[4] https://gemini.yahoo.com/advertiser/home.

[5] https://www.iab.com/guidelines/taxonomy/.

on Twitter[6] enables advertisers to target advertising campaigns on a community of users whose interests broadly align with the product or service to be promoted on Twitter. Interests are represented by 25 interest categories that expand into 350 sub-topics, ranging from "Education" to "Sports". For example, a web design company promoting an offer could choose to target "Technology" under the Business category, and "Web Design" under "Technology" to target users with those interests on Twitter.

Similarly, the "Interest Targeting" service on LinkedIn[7] also aims at letting the customers reach members with relevant ads that match their professional interests. For example, if an advertiser sells Cloud Computing services, the advertiser can target LinkedIn members interested in "cloud computing" and create a campaign with sponsored content ads showing how well the provided solution is helping business. Facebook and Instagram also provide "Detailed Targeting" option for advertisers so that they can define interests that targeted users should have, e.g., movies, books and TV.[8] Providing the option of choosing "interests" for advertisers in order to target users with those interests is common practise in other social media platforms such as Weibo[9] and Reddit.[10]

5.1.3 Personalized Content Summarization

Social media are effective user tools to create connections with other users and to cultivate and share personal interests. In general, users' connections and interests are the result of personal choices, operated while interacting with the platform (e.g., browsing timelines, following, searching, posting, replying, sharing contents, etc.). Due to the vastness of information flowing through social media platforms,[11] users are

[6] https://business.twitter.com/en/targeting/interest.html.

[7] https://business.linkedin.com/en-uk/marketing-solutions/blog/posts/marketing-solutions/2019/introducing-interest-targeting.

[8] https://www.facebook.com/business/help/440167386536513?id=176276233019487.

[9] https://www.dragonsocial.net/blog/chinese-social-media-weibo-and-twitter-comparison/.

[10] https://www.wordstream.com/blog/ws/2019/05/24/reddit-advertising.

[11] https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/.

literally flooded by tons of new contents everyday. Such availability of information is determining a growing dissatisfaction and is compromising the overall social media user experience.

Content summarization is the field of study which tries to cope with the problem of content quantity and content dimensionality. Works studying general purpose summarization techniques are focused on the type of media i.e., text, audio and video (Nazari and Mahdavi, 2019; Workie *et al.*, 2020) and try to cope with the problem of content dimensionality, offering a "smaller" version of the content that results more accessible to the majority of users. Example fields of applications for general purpose content summarization are: film industry, advertisement, recreation, video surveillance, data redundancy removal, computational time reduction, data visualization, search and retrieval.

Personalized content summarization, instead, is able to cope with both the problem of content quantity and content dimensionality. Content quantity and dimensionality are reduced to respect of user interest profiles, thus providing contents adapted to each individual social media user. For example, in Balan and Mathew (2020), the authors have experimented a significant and incremental growth of e-commerce sales when: (1) product descriptions are proposed in the original form, (2) descriptions are summarized with general purpose techniques, and finally (3) personalized summarization techniques are applied. The content quantity problem is also addressed in Ren *et al.* (2013) where the authors presented a tweets summarization approach and exploited a time-aware user behavior model, which is able to infer dynamic probabilistic distributions over interests and topics.

Personalized timeline summarization techniques can improve social media user experience. However, summarizing the timeline about an entity with social media data faces two main challenges: (1) key timeline episodes about the entity are unavailable in most social media platforms, and (2) since social media posts are short, noisy and informal, only content-based summarization is not sufficient. To address these challenges, in Chang *et al.* (2016), the authors have proposed a framework, named Timeline-Sumy, that utilizes episode detecting and summary ranking for personalized timeline summarization. Episode detecting is used to explicitly model temporal information and summary ranking is

used to rank social media posts in each episode via a learning-to-rank model. As another example, in order to facilitate user conversations about live streamed events, in Hannon *et al.* (2011a), the authors have proposed a Personalized Highlight Generator (PASSEV) that exploits user time-stamped opinions (mined from tweets) to generate video highlights for sport-related live events.

5.1.4 User Expertise Mining

As the numbers of users and contents are growing exponentially in social media, identifying experts in different topics is one of the useful applications that social media platforms can provide for their users to help them to find reliable information in different domains. The ability to understand the expertise of users is a key component to develop such applications (Xu, 2018). However, users are often unwilling to explicitly provide their expertise. User expertise modeling and user interest modeling can be considered as two related areas because both of them are trying to infer the relations between users and topics by processing the users' information in social media (e.g., the user's social posts, social relations, biographies and lists), such that the approaches applied for user expertise modeling mostly take advantage of user interest profiles.

There is a group of studies that seeks to model the *general topics of expertise* a user has knowledge of, based on her historical behavior (Guy *et al.*, 2013; Purohit *et al.*, 2012; Xu *et al.*, 2016, 2017). For example, in Ghosh *et al.* (2012), the authors have proposed utilizing the list memberships in Twitter to infer the topical expertise of users. They have used text processing techniques to analyse list metadata, and then, to infer the topics of expertise of the users in the list.

As some other examples, the authors in Xu *et al.* (2016) have inferred a user's expertise based on her posts on Twitter. They have first evaluated the importance of a tweet in identifying user expertise by using the sentiment intensity of the tweet. Then, an inference problem is modeled by exploiting the relatedness between expertise topics. In Wagner *et al.* (2012), the authors have learned the related topics of each user by running LDA over different types of user-related data

(e.g., tweets, retweets and lists). These topics are considered as a representation of the user's expertise. To evaluate the identified expertise and investigate the usefulness of each type of user data on Twitter for inferring user expertise, they have compared the identified expertise of a user with the user's actual expertise topics in the context of personalised recommendation.

In the literature, there is another group of studies that focuses on a particular *topic/domain of expertise* and tries to find experts in that domain or topic (Abbas *et al.*, 2015; Cheng *et al.*, 2014; Li *et al.*, 2014). For example, in Li *et al.* (2014), the problem of expertise modelling is investigated in terms of a user's knowledge on a place or a category of places instead of general topics/domains. The authors have analyzed the geo-tagged tweets of a user to identify if the user has knowledge about a given place. In their proposed model for user expertise mining on places, they have considered three properties of users: (1) within-topic activity (interactions the user had with the target place), (2) within-topic diversity (interactions the user had with other places in the category of the target place), and (3) recency (recent interactions the user had with the target place).

As other examples, in Abbas *et al.* (2015), the authors have proposed a method to find health experts from Twitter. Given a disease, they have first extracted the related keywords to the disease from the WordNet database.[12] Then, the initial expert candidates for the disease are selected based on the usage of these keywords in the user's posts. Finally, the final experts were selected for the given disease based on a metric defined from multiple criteria, such as the sentiment of the users in replies to the tweets of expert candidates. In Wei *et al.* (2016), the problem of expert finding on specific topics in Twitter was addressed as a search problem, in which each term in the query is treated as an expertise topic. The authors considered two aspects of users, i.e., global authority and local relevance, to estimate their level of expertise on a given topic. To model the global authority of a user on a topic, a semi-supervised graph-based probabilistic method was

[12] https://wordnet.princeton.edu/.

proposed that used three relations on Twitter (i.e., user-follower, user-list and list-list relations). The local relevance of users on a topic was calculated by a Gaussian-based method, where a user is represented by a document including the user's tweets, bio and the metadata of her list memberships.

5.1.5 Community Detection

Detecting communities on social media platforms such as Twitter or Facebook is useful to perform effective content and friend recommendations. In addition, it is also crucial in targeted advertising and viral marketing on social media platforms by identifying a subset of target users for advertising. Due to the importance of community detection on those social media platforms, many previous studies have focused on how user interests can effectively help to detect communities (Lim and Datta, 2012b).

For example, in Lim and Datta (2012a), the interest of a Twitter user in an interest category (or community) is inferred by the number of celebrities of that category that the user follows. The celebrities of each category are chosen based on the "occupation" of their corresponding Wikipedia concepts. Afterwards, communities are detected based on linkages among the followers of those celebrities. In Zhang *et al.* (2012), the authors have proposed discovering communities on Twitter based on the similarity between users, where the similarity incorporates different aspects with respect to a user: (1) the user's content including texts, URLs and hashtags in her tweets, and (2) the user's social behaviors such as following and retweeting. User interests based on texts here are inferred as the topics learnt by LDA considering the tweets of the user as a single document. Similarly, in Shi *et al.* (2017) user communities are detected based on the similarity between users, where user interest profiles are topic distributions inferred in a specific time interval.

A multimodal approach has been explored in Joshi *et al.* (2015) for community detection. The authors have exploited both shared photos and tags associated with those photos on Flickr to build user profiles and use them for detecting communities which shows better performance compared to considering only photos or tags. Recently,

in Fani *et al.* (2018) a graph-based approach is proposed to identify temporal topic-based user communities, i.e., communities of users who share similar topical interests with similar temporal behavior on Twitter. To model explicit interests of users over emerging topics on Twitter in different time intervals, they have utilized multivariate time series analysis. Then, they have built a user graph according to the measured user similarities based on the time series representation of the users. Finally, they have applied graph clustering techniques on the user graph to extract subgraphs that represent temporal user communities.

5.1.6 Aspect-Based Sentiment Analysis

The sentiment of user content and inferred user interest profiles has gained attention in several studies which has shown to be crucial for improving the recommendation performance in many domains such as e-commerce. For example, there is a line of studies (e.g., Li *et al.*, 2019; Wang *et al.*, 2018b; Zhang *et al.*, 2014) which has shown that considering the sentiment of user interests in terms of item features on e-commerce and review service websites, such as Amazon and Yelp, can improve the top-k recommendation performance and also provide explainable recommendation to the users.

In line with those efforts, several studies also have considered the sentiment of topics of (user) interest in the context of social media and its applications such as content or friend recommendations. For example, in Wang *et al.* (2018a), the authors have studied the problem of predicting the sentiment of user interests where the user interests refer to celebrities on Sina Weibo. The authors have claimed that the solution to this problem will be beneficial to many online services such as personalized advertising, new friends recommendation, public opinion analysis and opinion polls.

As another example, the authors in Yuan *et al.* (2014) have shown that considering sentiment homophily plays an important role for the friend recommendation task. The intuition behind their work is that, although two users both are interested in a certain topic (such as "Trump for President"), the two users may exhibit the same or contradictory sentiments toward the interest, which may play a crucial role in whether

they can be a friend or not. To this end, they have incorporated the sentiment toward the users' topics of interest, represented by hashtags on Twitter in their proposed approach. The results show that incorporating sentiment homophily as a set of features improves the performance of link prediction in terms of F1 score in both mutual-follow and mention graphs on Twitter.

The importance of considering the sentiment of user interests for content recommendations on Twitter has been studied as well. For example, the authors in Harakawa *et al.* (2018) have considered not only the topics included in a tweet (i.e., words, hashtags, and visual objects in the image of the tweet if any) that a user is interested in but also the sentiment regarding those topics in the context of recommending tweets that would be re-tweeted by a user. A multimodal Field-aware Factorization Machines (FFM) (Juan *et al.*, 2016) is used to incorporate those factors for sentiment-aware personalized tweet recommendations. The authors have shown that the use of sentiment of user interests in addition to other factors produces a higher MAP value.

5.2 Third-Party Applications

User interest modeling also enables the development of useful third-party applications beyond services on social media platforms. For example, third-party applications that allow *social login* functionality[13] can benefit from user interest modeling from social media to provide better services to their users. Social login functionality allows new users to an application to log-in using their social media accounts instead of creating new application specific accounts. Due to its growing popularity, many applications support the social login functionality so that users can log-in to those applications with their social media accounts such as Facebook or Twitter. Hence, this mechanism allows those applications, with the permission of the user, to access user-generated content and infer user interests to provide personalized services, even in cold-start situations (The integration of social media and third-party applications are discussed in more detail in Subsection 5.4).

[13] https://en.wikipedia.org/wiki/Social_login.

In the following, we provide an overview of some of the existing applications which exploit inferred user interests from various social media platforms for providing personalized services.

5.2.1 News Recommendations

One of the most popular applications taking advantage of user interest modeling on social media platforms, is recommending news articles, since the majority of the content generated on social media platforms (e.g., Twitter) is related to news (Kwak *et al.*, 2010). Indeed, social media have become the main source of news online, with more than two billion users, and nearly 64.5% of those users receive breaking news from Facebook, Twitter, YouTube, Snapchat and Instagram instead of traditional media.[14] In this regard, many previous studies have been proposed using inferred user interests on Twitter for personalized news recommendations.

For example, the authors in Phelan *et al.* (2009) have shown a content-based recommender system named Buzzer, which provides recommended real-time topical news articles from RSS news feeds using inferred user interest profiles from Twitter. A recommended article consists of the article content which includes a hyperlinked title of a news article from the RSS feed, and the summary of the article. In addition, each recommended news article has its metadata which consists of the recommendation score, and the associated terms of the article. Both user and article profiles are keyword-based and the weights of terms are measured by the TF-IDF score. Similarly, the authors in Abel *et al.* (2011b, 2013a) have used user interest profiles mined from users' tweets for recommending news articles. To this end, two types of keyword-based user profiles (based on hashtags and named entities) extracted from tweets are compared. In addition, the dynamics of user interests are incorporated into the user interest mining process by using temporal constraints (such as considering specific period(s)) or temporal patterns (e.g., weekday, night, etc.) which shows to be able to improve the quality of recommendation.

[14] https://www.forbes.com/sites/nicolemartin1/2018/11/30/how-social-media-has-changed-how-we-consume-news/#18a7813e3c3c.

Instead of using keyword-based approaches, in Zarrinkalam *et al.* (2015), the authors have used topics extracted in a given time interval to represent user interests. Here each topic is a set of semantic concepts (i.e., Wikipedia concepts) with high ephemeral correlations, i.e., the concepts have been frequently co-occurring together during a given interval. The resulting user interest profiles, built for each time interval, are then used for recommending news from BBC or CNN during that time interval, which shows better performance compared to those interest mining approaches used in Abel *et al.* (2013a).

5.2.2 Point-of-Interest (POI) Recommendations

A point-of-interest (POI) is a specific type of interest with respect to locations. Many social media platforms, such as Twitter and Flickr, provide the geo-tagging functionality to users, so that they can generate content with geo-tags, which are metadata associated to a user's geographical location. Mining geo-tagged user-generated contents and inferring related locations enables the creation of useful applications such as POI recommender systems and personalized tour planners.

For example, the authors in Lim (2015) have proposed an approach for recommending personalized tours based on user interests from her visit history on Flickr (consisting of a collection of geo-tagged photos). To this end, POI alignments to Wikipedia concepts, and the related entities have been used for inferring user interests. For example, https://en.wikipedia.org/wiki/Category:Landmarks_in_Melbourne is a POI category which contains dozens of entities such as https://en.wikipedia.org/wiki/Bolte_Bridge and https://en.wikipedia.org/wiki/Luna_Park,_Melbourne. Flickr photos shared by users are mapped to the extracted list of POIs based on their coordinates (provided in the form of geo-tags). Then, the user interest profiles are used for recommending personalized tours (as a sequence of POIs). Similarly, in Lim *et al.* (2018), the authors have proposed a personalized tours recommender system based on POIs of users on Twitter. Different from Lim (2015), user interests are represented as POI categories where the weights of each category is determined by her time spent at and the number of times the user visits each POI in the category.

As another example, the authors in Abel *et al.* (2012) have proposed recommending POIs for users based on their activities on Twitter and Flickr. Concept-based user interest profiles are built based on the users' posts on Twitter and their sharing activities on Flickr. In this work, each concept represents a DBpedia entity related to a POI such as http://dbpedia.org/page/Eiffel_Tower. Each mined concept is further enriched based on its indirect links in the DBpedia graph. For example, http://dbpedia.org/page/Paris will be added into a user's interest profiles if http://dbpedia.org/page/Eiffel_Tower is one of her explicit interests since they are indirectly linked via the property http://dbpedia.org/ontology/location in the DBpedia knowledge graph.

5.2.3 Research-Related Recommendations

Social media platforms such as Twitter have also been widely used by researchers for sharing/announcing scientific related events and information (Letierce *et al.*, 2010), and it has been shown that professional interests extracted from the social media activities of researchers are highly similar to the ones extracted from their publications (Große-Bölting *et al.*, 2015). This motivates and enables many interesting applications of user interests mining with respect to researchers based on their online social media activities as well.

For instance, an application to recommend researchers is proposed in Nishioka *et al.* (2015) based on the user interest profiles inferred by leveraging users research related Twitter activities. To this end, domain-specific concepts explicitly mentioned in the tweets of researchers/users are extracted to build concept-based user interest profiles. For example, the ACM Computer Classification System (CCS)[15] is used as a hierarchical knowledge base for the computer science domain, and the Medical Subject Headings (MeSH)[16] is used as the one for the domain of medicine. Furthermore, implicit user interests (i.e., concepts that are not mentioned explicitly in the text) are identified by exploiting the hierarchical structure of those knowledge bases related to the explicitly mentioned concepts. Afterwards, inferred professional interests

[15] https://www.acm.org/publications/class-2012.
[16] https://www.nlm.nih.gov/databases/download/mesh.html.

of users are used for recommending researchers that a user might be interested in.

In addition, several studies have presented the application of recommending research publications using inferred interest profiles from Twitter. In Nishioka and Scherp (2016), the authors have used similar user interest mining approaches proposed in Nishioka *et al.* (2015) to infer professional interests in the domain of economics, i.e., user interests are represented as concepts in a hierarchical knowledge base with respect to the domain. Different from Nishioka *et al.* (2015), the authors have also leveraged interest decay functions computed, for example by combining a sliding time window and exponential decay function to give higher importance to the concepts recently mentioned on user's tweets. The authors showed that using inferred user interest profiles with the titles of scientific publications are sufficient to achieve competitive recommendation results when employing the proposed interest mining approach.

5.3 Other Applications

In the following, we provide an overview of other interesting applications using inferred interest profiles based on various user interest mining approaches introduced in this work. Estimating the location of a user when she is generating contents in social media, e.g., where a user tweeted about some events, is important to many *smart city* applications such as understanding the mobility patterns of users in a city. Inferred user interests can be used for estimating a user's location correctly. For example, in Chen *et al.* (2013), the authors have used interest profiles inferred from Chinese tweets obtained from Sina Weibo to predict user locations. To this end, the authors represent user interest profiles as topic distributions where the topics are learned using LDA based on their tweets. A mapping between location function (e.g., sports or entertainment) and user interest is established first and used for estimating actual locations.

Recently, the authors in Inaba and Takahashi (2018) have proposed inferring user interests from open-domain chat-oriented dialogue systems in order to provide personalized responses. This allows the system to

enhance both user engagement and satisfaction by focusing on topics a user is interested in and avoiding topics they dislike. User interests are represented with 24 topic categories (out of the categories used by Yahoo Chiebukuro[17]) which were likely to be related to the content appearing in one-to-one dialogues between strangers. The authors used a three-point scale to represent the degree of interests in each topic i.e., 0 (disinterest, indifference and neutrality), 1 (light interest), or 2 (strong interest). The degree of interest is estimated by a neural network-based approach proposed by the authors using utterances in chat dialogues.

In Lully *et al.* (2018), the authors have shown that user interest profiles extracted from interactions between users and photos can be used for recommending tour-related photos to be displayed in ads banners. To this end, concept-based user interest profiles are extracted based on interacted photos of a user with the following steps. First, each photo is classified into the 1000 ImageNet concepts using a computer vision method named "Inception-V3" (Szegedy *et al.*, 2016). Concepts are then mapped into corresponding DBpedia entities with a semi-automatic approach. For each photo, other photos that are highly related to the given photo can be retrieved based on their similarities using the embeddings of those photos learned by Inception-V3. Then, user interest profiles are enriched, with implicit user interests by mining those concepts related to explicit interests, from the DBpedia knowledge graph. Although the interest mining approach is based on photos that a user has interacted with in a system of travel agency, the same approach can be used for inferring user interests based on the photos shared by users in social media platforms (such as Instagram or Flickr) as well, which can be useful for those cold-start users who have not interacted with many photos in the system.

The abundant user-generated content (such as texts and photos) on many existing social media platforms provides a great opportunity for many start-ups to better understand user interests and provide personalized services. For example, a start-up company which sells and recommends movies can provide personalized movie recommendations based on inferred user interest profiles from their social media.

[17] https://chiebukuro.yahoo.co.jp/.

Digi.me,[18] which enables individual users to aggregate and manage their social media data at a central place, is a good example of providing third-party applications rich user profiles, including their interests for improving their services (with the permission of users) such as personalized e-commerce. In this regard, we expect that there will be a growing number of applications of user interest profiles in the future.

5.4 Integration of Social Media and Third-Party Applications

In recent years, we have observed a significant growth in the integration of third-party applications with social media platforms. Because personal user information (such as user interests, user generated content and user social relationships) from social media can be leveraged with the permission of a user to provide advanced functionalities (e.g., personalized services and personalized advertising) for third-party applications.

In order to be able to integrate third-party applications with social media platforms, most social media offers services to allow third-party applications to programmatically interact with these platforms. Due to the "online" nature of social media, these services are deployed in the form of representational state transfer (REST) application programming interface (API) which provides the access to social media functionalities through hypertext transfer protocol (HTTP) requests.[19] In the following, we introduce some important examples of such services:

1. Social login: This is a functionality that allows users to log-in by using their social media accounts instead of creating new application specific accounts.

2. Accessing user related data: third-party applications are able to ask an explicit user agreement to access her own personal data. Depending on the social media platform, available personal user

[18] https://digi.me/.

[19] Twitter: https://developer.twitter.com/en.
Facebook: https://developers.facebook.com/docs/apis-and-sdks.
Weibo: https://open.weibo.com/wiki/SDK/en.
Instagram: https://www.instagram.com/developer/.
Pinterest: https://developers.pinterest.com/docs/getting-started/introduction/.

information include: status updates, lists of followers and followees, subscriptions, short bio, email, age, gender and location.

3. Performing status updates: with the agreement of the user, third-party applications can generate user status updates typically to share application generated contents.

4. Establishing social relationships: third-party applications can offer functionalities to allow users to follow (or to subscribe) a specific social media account (or channel) without using social media platform dedicated user interfaces. For example, a website may offer widgets to let a user start following the status updates of its social accounts (i.e., Facebook, Twitter and LinkedIn).

In general, thanks to REST API, all kinds of applications[20] – when a functioning internet connection is available – can interface with social media platforms and as a consequence access to personal end-user data. A notable example of integration is the popular video streaming and social media platform – YouTube. YouTube is deployed in many forms (e.g., as a web application or a mobile application) and provides immediate tools to perform the action of sharing video links in status updates on a user's social media. In addition, some social media platforms (e.g., Facebook) are offering more sophisticated API enabling the integration of third-party applications within the platform itself.

Based on REST API, third-party applications can be designed to collect personal information for a large population of users. It is worth noting that massive personal user data collection activities are performed in different fields of application, ranging from research activities (e.g., large-scale datasets are required for training and evaluation purposes see Section 4) to commercial activities (e.g., marketing and advertising purposes).

To integrate third-party applications with social media platforms, the systematic access and management of personal user data has been a widely discussed topic. To protect the privacy of their users, social media platforms adopted more and more restrictive rules on the types

[20] A social media platform can behave as a third-party application and interact with other social platforms.

and quantities of user data accessible by third-party applications. The privacy policies of a social media platform are clearly stated and must be accepted (with explicit agreement) by all the developers of third-party applications who are willing to use the platform REST API. Additionally, social media platforms adopted technological mechanisms to monitor and to prevent any possible violation. Moreover, some institutions and governments regulated the management of personal user data (thus involving social media platforms) with advanced legislative tools. Notable examples are the General Data Protection Regulation (GDPR)[21] and the California Consumer Privacy Act (CCPA).[22]

Despite the fact that the problem is widely felt globally, there are still many important aspects related to personal user data and privacy that require regulations to be globally elaborated. As an example, the "right to be forgotten" – the right to have negative personal information removed from Internet searches and other directories under certain circumstances – has been discussed and put into practice in the European Union and Argentina only.[23] More information on the subject can be found in Bakhoum *et al.* (2018) where different authors analyse the legal approach to personal data taken by different fields of law.

[21] https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[22] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

[23] https://en.wikipedia.org/wiki/Right_to_be_forgotten.

6

Open Challenges and Future Directions

With respect to the exciting open research questions in the state-of-the-art for mining users' interests from online social media, this section will present different challenges and future directions which are related to: (1) semantics, (2) cross-system models, (3) dynamicity, (4) comprehensiveness, (5) explainability, and (6) reproducibility.

6.1 Semantics

Some studies (see Section 3) have proposed to enrich social posts by annotating them with unambiguous concepts described in external knowledge bases such as Wikipedia/DBpedia. In other words, entity linking is a fundamental step in *concept-based* user interest modeling. However, since social posts are rather short, noisy and informal and they often do not provide sufficient contextual information, entity linking from social posts is challenging by itself. For doing this step, most studies rely only on the existing entity annotators such as TagMe, Zemanta or Aylien. Although the accuracy of this step has influence on the accuracy of the inferred interests, many studies don't consider the uncertainty of these annotators. Therefore, more investigations on the impact of the

accuracy of annotators on the performance of user interest modeling are needed.

Furthermore, existing studies only consider explicit entities which are mentioned in the social posts to infer user interests. For example, in the following tweet, the term *Gravity* is an explicit entity mentioned on the tweet:

> The movie Gravity was more expensive than the Mars Orbiter Mission.

Concept-based user interest modeling methodologies use explicit entities to model user interests. However, there are some entities in the social media posts which are implicitly mentioned. For example, in the following tweet, the term *Gravity* is not explicitly mentioned in the tweet, even if however the tweet is talking about this movie:

> this new space movie is crazy, you must watch it.

Implicit entity linking has recently been introduced to link a post to a Wikipedia entity while the name of that entity or its synonyms are not explicitly mentioned in the post. In the first work that introduced this concept (Hosseini *et al.*, 2018), the authors have shown the importance of implicit entity linking. It is mentioned that 21% of the entities in the Movie domain and 40% in the book domain are implicit entities. Therefore, applying implicit entity linking methods for extracting implicit entities from posts and infer interests of users more accurately is an interesting future direction which is yet to be explored.

6.2 Cross-System Models

The majority of existing approaches of user interest modeling from social media have mainly focused on a single social media as a source of information. Most of these studies use Twitter because the information that the users publish are more publicly accessible compared to other social media. However, people expose and share different aspects of their behavior, personality and activities in different social media. According to GlobalWebIndex,[1] on average, people actively use 2.8 social media

[1] https://www.globalwebindex.com/.

platforms and this value is increasing over the years. Therefore, to provide a comprehensive view of user's behaviors, it is crucial to integrate information from multiple social media.

Despite its advantages, this leads to important data integration challenges: entity matching and duplicates resolution, conflict resolution, heterogeneity of the data models of the sources and the consequent need of a common target data model, to name a few (Orlandi *et al.*, 2012). New technologies have been recently developed to link different social media accounts of the same user (Shu *et al.*, 2016). Thus, as a future direction in user interest modeling from social media, with linked data, it is possible to perform novel cross-system user interest modeling such as: (1) domain bias analysis that focuses on studying the impact of domain or social network platform on user interests, and (2) domain adaptation that addresses how to adapt prediction models trained on one platform (e.g., Twitter) to predict the user interests on other platforms (e.g., Facebook) (Pan and Ding, 2019).

6.3 Dynamicity

In terms of the dynamicity dimension in user interest modeling from social media, it has been well accepted that users' degree of interest changes over time and is considered in different studies. However, social media have more dynamicity in other aspects. For example, new users may join a social network or current users may leave; new topics may emerge or current topics may disappear; even the relationship between users or topics change over time on a social network. Thus, more dynamicity should be considered in user interest modeling approaches. In Zarrinkalam *et al.* (2018) the authors have already shown that modeling social network information as a unified graph and applying heterogeneous link prediction methods is promising for modeling user interests. However, they have overlooked the dynamicity of information in social media. Recently, in Fard *et al.* (2019) the authors have proposed an approach for link prediction in dynamic heterogeneous information networks which is shown that works well in the context of co-author prediction. Applying approaches that consider both heterogeneity and

dynamicity of information in the context of interest modeling can be considered as a future direction.

By modeling the information of social media as a dynamic graph for predicting user's interests, a promising approach to infer interests of a user is applying dynamic network embedding methods. Dynamic network embedding has shown promising results in underlying tasks such as node classification (Cavallari *et al.*, 2017), link prediction (Grover and Leskovec, 2016) and community detection (Perozzi *et al.*, 2014), among others. The problem with the majority of existing dynamic network embedding methods is that they overlook the heterogeneity of a graph. Recently, in Bian *et al.* (2019) the authors have proposed a graph embedding method that considers both heterogeneity and dynamicity of the underlying graph to learn the vector representations of nodes. Applying such approaches in the context of interest modeling can be considered as a future direction. Recently, in Chen *et al.* (2020), the authors have proposed a recommender system which is able to embed user item ratings with temporal information. The proposed methodology is based on graph embeddings and diffusion sampling (graph2vec) and embeds user's behavior histories to a low-dimensional vector space.

6.4 Comprehensiveness

Investigating the synergistic effect of different dimensions is important for developing better user modeling methods. For example, in the context of link recommendation application, in Piao and Breslin (2016d) the authors investigated the effect of four dimensions in user interest modeling, i.e., interest representation, enrichment, temporality of user interests, and interest propagation. They have shown that representing user interests using WordNet synsets and DBpedia entities and enriching posts by the content of the embedded URLs in posts is an effective approach in the context of link recommendations on Twitter. This study can be considered as a good starting point for more comprehensive user interest modeling methods. However, in their work, the effect of different dimensions is only investigated in a specific context, i.e., link recommendation.

Different user interest modeling strategies perform differently in different applications. For example, in the context of news recommendation, the authors in Abel *et al.* (2011b) have shown that using interest decay function for constructing user interest profiles on Twitter performs better than applying constraint-based approaches, which is different from the results reported by the authors of Nishioka and Scherp (2016) in the context of publication recommendations. Therefore, more comprehensive studies are needed to answer the important research question "Which combinations of different approaches in each dimension can provide the best user interest profiles in each application?".

6.5 Explainability

The explainability of recommender systems has attracted considerable attention by researchers to: (1) make recommendation models better understandable, i.e., to help users to understand how the system works (Transparency), and (2) allow users to tell the system if it is wrong (Scrutability) (Balog *et al.*, 2019). The research on explainable recommendation methods has already been investigated in different application domains, such as explainable e-commerce recommendation (He *et al.*, 2015c; Zhang *et al.*, 2014), explainable social recommendation (Ren *et al.*, 2017), and explainable point-of-interest recommendation (Wang *et al.*, 2018c). Interested readers are encouraged to see Zhang and Chen (2020) for the state of the art on explainable recommendation.

However, explainability as an important dimension in user interest modeling from online social media is overlooked. Building user models in a manner that engages the user in a feedback cycle may improve the quality of the model and the user's control over the personalization. Recently, in the context of movie recommendation, the authors in Balog *et al.* (2019) have taken the explainability to the level of user interests. Instead of explaining to the user why a given item was recommended, they have presented an approach to provide a textual description that summarizes the system's understanding of the user's preferences. They have also allowed the user to scrutinize this summary and thereby directly modify the user model. Therefore, incorporating explainability as a dimension into user modeling strategies in online social media

is a promising future direction and deserves further investigation and evaluation.

6.6 Reproducibility

Due to the lack of common benchmarks and datasets in the context of user interest modeling from online social media, although there are many studies in this field, they don't compare their proposed approach with previous studies. Recently, in Tommaso *et al.* (2018) the authors have presented a user interests dataset which seems promising for comparing the approaches in this context (see Section 4). It includes half a million Twitter users with an average of 90 multi-domain preferences per user on music, books, etc. To reliably extract a user's interests, they have used two sources of information: (1) those tweets of the user which are published via online platforms such as Spotify, Goodreads and IMDB, and (2) the user's friendship list.

Another challenge of reproducibility in user interest modeling from social media is that the implementation of current studies are not available. Therefore, it is hard for researchers to reimplement the approaches proposed in previous studies for comparison. As a future direction, providing a user interest modeling library that include the implementation of the state of the art approaches, in the same way as Librec[2] library (Guo *et al.*, 2015) for recommender systems is important to evaluate the performance of different user interest modeling strategies.

[2] https://github.com/guoguibing/librec.

Appendices

A.1 Glossary

A

Annotation is extra information associated with a particular point in a document or other piece of information.[3]

B

Big Five Model is a suggested taxonomy, or grouping, for personality traits, developed from the 1980s onwards in psychological trait theory. The Big Five personality traits is also known as the five-factor model (FFM) and the OCEAN model.[4]

Biography/Bio (Social Media) refers to a short bit of explainer text that explains who the user is.[5]

C

Category (Wikipedia) is intended to group together Wikipedia articles on similar subjects. Clicking a category name brings up a category page listing the articles (or other pages) that have been added to that particular category.[6]

Co-Occurrence Graph/Network is the collective interconnection of terms based on their paired presence within a specified unit of text. For example, terms A and B may be said to "co-occur" if they both appear in a particular article.[7]

Cross-System User Modeling is a user modelling methodology that interweaves user profiles from diverse Social Media (Abel *et al.*, 2013b).

[3] https://en.wikipedia.org/wiki/Annotation.
[4] https://en.wikipedia.org/wiki/Big_Five_personality_traits.
[5] https://blog.hubspot.com/marketing/social-media-terms.
[6] https://en.wikipedia.org/wiki/Help:Category.
[7] https://en.wikipedia.org/wiki/Co-occurrence_network.

D

Decay Function describes the process of reducing an amount by a consistent percentage rate over a period of time.[8]

E

Entity Linking is the task of assigning a unique identity to entities (such as famous individuals, locations, or companies) mentioned in text. In natural language processing, entity linking is also referred to as named-entity linking (NEL), named-entity disambiguation (NED), named-entity recognition and disambiguation (NERD) or named-entity normalization (NEN).[9]

Explicit Interest is an interest of a user extracted by leveraging information from the user's own activities (e.g., social posts and social relations).

Extrinsic Evaluation assesses the performance of a text mining system (user modeling in our context) component from the perspective of its effects to the performance of the whole system (Suominen, 2009).

F

Followee (Social Media) refers to a user B who is followed by another user A. The user A is able to receive status updates of user B.

Follower (Social Media) refers to a user B who subscribes to the user A account in order to receive status updates of user A.

G

Geotag refers to a piece of electronic data that shows where someone or something is and can, for example, be attached to a photograph or comment on social media.[10]

[8] https://www.thoughtco.com/exponential-decay-definition-2312215.
[9] https://en.wikipedia.org/wiki/Entity_linking.
[10] https://dictionary.cambridge.org/dictionary/english/geotag.

Google Knowledge Graph is a knowledge base used by Google and its services to enhance its search engine's results with information gathered from a variety of sources.[11]

Ground Truth is a term used in various fields to refer to information provided by direct observation (i.e., empirical evidence) as opposed to information provided by inference.[12]

H

Hashtag is a tag used on a variety of social media as a way to annotate a message. A hashtag is a word or phrase preceded by a # (e.g., #InboundMarketing). Social media users use hashtags to categorize information and make it easily searchable.[13]

Homophily refers to the tendency of people to have (non-negative) ties with people who are similar to themselves in socially significant ways.[14]

Human Readable A human-readable medium or human-readable format is any encoding of data or information that can be naturally read by humans.[15]

I

Implicit Interest is a potential interest that a user did not explicitly mention but might have an interest in. Relationship between users and relationship between topics of interest are two indicators to infer implicit interests of users (Zarrinkalam *et al.*, 2016).

[11] https://en.wikipedia.org/wiki/Knowledge_Graph.
[12] https://en.wikipedia.org/wiki/Ground_truth.
[13] https://blog.hubspot.com/marketing/social-media-terms.
[14] http://www.analytictech.com/mgt780/topics/homophily.htm.
[15] https://en.wikipedia.org/wiki/Human-readable_medium.

Infobox a consistently-formatted table which is present in articles with a common subject.[16]

Interoperability is a characteristic of a product or system, whose interfaces are completely understood, to work with other products or systems, at present or in the future, in either implementation or access, without any restrictions.[17]

Intrinsic Evaluation assesses the performance of a text mining system (user modeling in our context) component as an isolated unit unconnected to the other system components (Suominen, 2009).

K

Knowledge Base is a centralized database for spreading information and data. Knowledge bases support collecting, organizing, retrieving, and sharing knowledge.[18]

Knowledge Graph in general, any graph-based representation of some knowledge could be considered a knowledge graph. In many fields of science a knowledge graph (Paulheim, 2016): (1) mainly describes real world entities and their interrelations, organized in a graph, (2) defines possible classes and relations of entities in a schema, (3) allows for potentially interrelating arbitrary entities with each other, (4) covers various topical domains.

L

Latent User Modeling refers to user modelling techniques devoted to the definition of user interests models built on top of (or just including) latent features.

[16] https://en.wiktionary.org/wiki/infobox.
[17] https://en.wikipedia.org/wiki/Interoperability.
[18] https://blog.hubspot.com/service/what-is-a-knowledge-base.

Like (Social Media) Likes are actions that can be made by some social media users (e.g., Facebook, Instagram and Twitter users). Instead of writing a comment or sharing a post, a user click the Like button as a quick way to show approval.[19]

List Membership (Social Media) (see Topical List).

O

Ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application (Gruber, 2009).

P

Pin (Social Media) is the basic unit of Pinterest. It is the equivalent to the tweet on twitter. Pins consist of an image or video. Pins also include information specifying the board in which it is pinned, the source (either uploaded or taken from the web) and an indication from whom it was repined (if indeed it was).[20]

Post (Social Media) a piece of writing, image, or other item of content published online, typically on a blog or social media website or application.[21]

R

RDF Schema (Resource Description Framework Schema, variously abbreviated as RDFS, RDF(S), RDF-S, or RDF/S) is a set of classes with

[19] https://blog.hubspot.com/marketing/social-media-terms.
[20] https://corp.wishpond.com/pinterest-marketing-resources-for-business/fundamentals-of-pinterest-pins-boards-repinning-and-everything-inbetween/.
[21] https://www.lexico.com/definition/post.

certain properties using the RDF extensible knowledge representation data model, providing basic elements for the description of ontologies, otherwise called RDF vocabularies, intended to structure RDF resources. These resources can be saved in a triplestore to reach them with the query language SPARQL.[22]

Recommender System a subclass of information filtering system that seeks to predict the rating and preference a user would give to an item (Khanian Najafabadi *et al.*, 2017).

Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax notations and data serialization formats. It is also used in knowledge management applications.[23]

Retweet (Social Media) is a functionality to repost (something, such as a message) to the Twitter online message service.[24]

S

Semantic Web is an extension of the World Wide Web through standards set by the World Wide Web Consortium (W3C). The goal of the Semantic Web is to make Internet data machine-readable. To enable the encoding of semantics with the data, technologies such as Resource Description Framework (RDF) and Web Ontology Language (OWL) are used. These technologies are used to formally represent metadata.[25]

Skip-Gram is one of the unsupervised learning techniques used to find the most related words for a given word. Skip-gram is used to predict the context word for a given target word. It's the reverse of CBOW

[22] https://en.wikipedia.org/wiki/RDF_Schema.
[23] https://en.wikipedia.org/wiki/Resource_Description_Framework.
[24] https://www.merriam-webster.com/dictionary/retweet.
[25] https://en.wikipedia.org/wiki/Semantic_Web.

algorithm. Here, target word is input while context words are output. As there is more than one context word to be predicted which makes this problem difficult.[26]

Social Media is a Web 2.0 applications, where: (1) user-generated content is the lifeblood, (2) individuals and groups create user-specific profiles for a site or app designed and maintained by a social media service, (3) social media services facilitate the development of social networks online by connecting a profile with those of other individuals and/or groups (Obar and Wildman, 2015).

Social Network is a network of individuals (such as friends, acquaintances, and coworkers) connected by interpersonal relationships. In the context of social media a social network is the network of social relationships established between users.[27]

SPARQL (pronounced "sparkle", a recursive acronym for SPARQL Protocol and RDF Query Language) is an RDF query language - that is, a semantic query language for databases-able to retrieve and manipulate data stored in Resource Description Framework (RDF) format. It was made a standard by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium, and is recognized as one of the key technologies of the semantic web.[28]

Status Update is an update feature (Facebook) which allows users to discuss their thoughts, whereabouts, or important information with their friends. Similar to a tweet on the social networking site Twitter, a status is usually short and generally gives information without going into too much detail.[29]

[26] https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c.
[27] https://www.merriam-webster.com/dictionary/socia_network.
[28] https://en.wikipedia.org/wiki/SPARQL.
[29] https://whatis.techtarget.com/definition/Facebook-status.

T

Tag (Social Media) is a label that allows users to create a link back to the profile of the person shown in the picture or targeted by the update.[30]

Taxonomy is the science of classification according to a predetermined system, with the resulting catalog used to provide a conceptual framework for discussion, analysis or information retrieval. In theory, the development of a good taxonomic classification takes into account the importance of separating elements of a group (taxon) into subgroups (taxa) that are mutually exclusive and unambiguous, and taken together, include all possibilities.[31]

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Rajaraman and Ullman, 2011).

Timeline (Social Media) is a user collection of images, and tagged posts kept in reverse chronological order. Timeline is often part of social media user profile pages to provide a view of recent user activities.

Topical Lists/List Membership (Social Media) is a curated list/group of Twitter accounts. One can create his own Lists or subscribe to Lists created by others users. Viewing a List timeline will show you a stream of Tweets from only the accounts on that List.[32]

U

User Identity Linkage is aimed at predicting if two or more user profiles across different Online Social Media platforms are related to the same user identity (Shu *et al.*, 2016).

[30] https://blog.hubspot.com/marketing/social-media-terms.
[31] https://searchcontentmanagement.techtarget.com/definition/taxonomy.
[32] https://help.twitter.com/en/using-twitter/twitter-lists.

User Interest Modeling is the process of obtaining the user interest profile (Piao and Breslin, 2018a).

User Interest Profile is a data structure that represents the degree of interest of an individual user over a set of topics which are human readable (e.g., represented by words or concepts).

User Study Evaluation assesses the performance of a system with real end users, generally conducted in usability laboratories under controlled settings (Lumsden, 2008).

A.2 Acronyms

AI Artificial Intelligence
API Application Programming Interface
AUC Area Under the Curve
CCPA California Consumer Privacy Act
dDTM discrete Dynamic Topic Model
FOAF Friend Of A Friend
FPM Frequent Pattern Mining
GRP General Data Protection Regulation
HTTP Hypertext Transfer Protocol
LDA Latent Dirichlet Allocation
MAP Mean Average Precision
ML Machine Learning
MRR Mean Reciprocal Rank
NDCG Normalized Discounted Cumulative Gain
OPM Open Provenance Model
P@N Precision at rank N
POI Point of interest
R@N Recall at rank N
RDF Resource Description Framework
REST Representational State Transfer
RSS RDF Site Summary
S@N Success at rank N
SPARQL SPARQL Protocol and RDF Query Language

SCOT Social Semantic Cloud of Tags
SIOC Semantically Interlinked Online Community
SKOS Simple Knowledge Organization System
SW Semantic Web
URL Uniform Resource Locator
WI Weighted Interests
WO Weighting Ontology
XML Extensible Markup Language

References

Abbas, A., M. U. S. Khan, M. Ali, S. U. Khan, and L. T. Yang (2015). "A cloud based framework for identification of influential health experts from Twitter". In: *2015 IEEE 12th Intl. Conf. on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl. Conf. on Autonomic and Trusted Computing and 2015 IEEE 15th Intl. Conf. on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), Beijing, China, August 10–14*. 831–838.

Abel, F., N. Henze, E. Herder, and D. Krause (2010). "Linkage, aggregation, alignment and enrichment of public user profiles with mypes". In: *Proceedings of the 6th International Conference on Semantic Systems. I-SEMANTICS '10*. Graz, Austria: ACM. 11:1–11:8.

Abel, F., Q. Gao, G. Houben, and K. Tao (2011a). "Analyzing temporal dynamics in Twitter profiles for personalized recommendations in the social web". In: *WebSci '11: Proceedings of the 3rd International Web Science Conference, Koblenz, Germany, June 15–17*. Article 2, 1–8.

Abel, F., Q. Gao, G. Houben, and K. Tao (2011b). "Analyzing user modeling on Twitter for personalized news recommendations". In: *User Modeling, Adaption and Personalization – 19th International Conference, UMAP 2011, Girona, Spain, July 11–15. Proceedings*. 1–12.

Abel, F., Q. Gao, G. Houben, and K. Tao (2011c). "Semantic enrichment of Twitter posts for user profile construction on the social web". In: *The Semanic Web: Research and Applications – 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29–June 2, Proceedings, Part II*. 375–389.

Abel, F., C. Hauff, G. Houben, and K. Tao (2012). "Leveraging user modeling on the social web with linked data". In: *Web Engineering – 12th International Conference, ICWE 2012, Berlin, Germany, July 23–27. Proceedings*. 378–385.

Abel, F., Q. Gao, G.-J. Houben, and K. Tao (2013a). "Twitter-based user modeling for news recommendations". In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. IJCAI '13*. Beijing, China: AAAI Press. 2962–2966. URL: http://dl.acm.org/citation.cfm?id=2540128.2540558.

Abel, F., E. Herder, G. Houben, N. Henze, and D. Krause (2013b). "Cross-system user modeling and personalization on the social web". *User Model. User-Adapt. Interact.* 23(2–3): 169–209.

Adomavicius, G. and A. Tuzhilin (2005). "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions". *IEEE Trans. Knowl. Data Eng.* 17(6): 734–749.

Ahmed, A., Y. Low, M. Aly, V. Josifovski, and A. J. Smola (2011). "Scalable distributed inference of dynamic user interests for behavioral targeting". In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21–24*. 114–122.

Aiello, L. M., G. Petkos, C. J. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes (2013). "Sensing trending topics in Twitter". *IEEE Trans. Multimedia*. 15(6): 1268–1282.

Alvarez-Melis, D. and M. Saveski (2016). "Topic modeling in Twitter: Aggregating tweets by conversations". In: *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17–20*. 519–522. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13162.

Anita, W. (2013). "Why people use social media: A uses and gratifications approach". *Qualitative Market Research*. 16(4): 362–369.

Antelmi, A., D. Malandrino, and V. Scarano (2019). "Characterizing the behavioral evolution of Twitter users and the truth behind the 90-9-1 rule". In: *Companion of the 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13–17*. 1035–1038.

Arabzadeh, N., H. Fani, F. Zarrinkalam, A. Navivala, and E. Bagheri (2018). "Causal dependencies for future interest prediction on Twitter". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22–26*. 1511–1514.

Asur, S. and B. A. Huberman (2010). "Predicting the future with social media". In: *WI2010*. 492–499.

Bakhoum, M., B. C. Gallego, M.-O. Mackenrodt, and G. Surblyt-Namaviien (2018). *Personal Data in Competition, Consumer Protection and Intellectual Property Law: Towards a Holistic Approach?* 1st edn. Springer.

Balan, M. and S. Mathew (2020). "Personalize, summarize or let them read? A study on online word of mouth strategies and consumer decision process". *Information Systems Frontiers*. DOI: 10.1007/s10796-020-09980-9.

Balog, K., F. Radlinski, and S. Arakelyan (2019). "Transparent, scrutable and explainable user models for personalized recommendation". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21–25*. 265–274.

Bao, H., Q. Li, S. S. Liao, S. Song, and H. Gao (2013). "A new temporal and social PMF-based method to predict users' interests in microblogging". *Decision Support Systems*. 55(3): 698–709.

Benton, A., R. Arora, and M. Dredze (2016). "Learning multiview embeddings of Twitter users". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, Berlin, Germany, Volume 2: Short Papers*. URL: https://www.aclweb.org/anthology/P16-2003/.

Berners-Lee, T. (2005). "WWW at 15 years: Looking forward". In: *Proceedings of the 14th International Conference on World Wide Web. WWW '05*. Chiba, Japan: ACM. 1–1.

Besel, C., J. Schlötterer, and M. Granitzer (2016). "Inferring semantic interest profiles from Twitter followees: Does Twitter know better than your friends?" In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, April 4–8*. 1152–1157.

Bhardwaj, A., A. Blarer, P. Cudre-Mauroux, V. Lenders, B. Motik, A. Tanner, and A. Tonon (2019). "Event detection on microposts: A comparison of four approaches". *IEEE Transactions on Knowledge and Data Engineering*. PP(Oct.): 1–1.

Bhattacharya, P., M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi (2014). "Inferring user interests in the Twitter social network". In: *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA – October 6–10*. 357–360.

Bian, R., Y. S. Koh, G. Dobbie, and A. Divoli (2019). "Network embedding and change modeling in dynamic heterogeneous networks". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21–25*. 861–864.

Blei, D. M. (2012). "Probabilistic topic models". *Commun. ACM*. 55(4): 77–84.

Blei, D. and J. Lafferty (2006). "Dynamic topic models". In: *ICML '06 – Proceedings of the 23rd International Conference on Machine Learning*. 113–120.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). "Latent dirichlet allocation". *Journal of Machine Learning Research*. 3: 993–1022. URL: http://www.jmlr.org/papers/v3/blei03a.html.

Bojārs, U., J. Breslin, A. Finn, and S. Decker (2008). "Using the semantic web for linking and reusing data across Web 2.0 communities". *Journal of Web Semantics*. 6(1): 21–28.

Bojars, U., A. Passant, J. Breslin, and S. Decker (2008). "Social networks and data portability using semantic web technologies". In: *2nd Workshop on Social Aspects of the Web (SAW 2008), in Conjunction with 11th International Conference on Business Information Systems (BIS 2008)*.

Boldi, P. and C. Monti (2016). "Cleansing Wikipedia categories using centrality". In: *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11–15, Companion Volume*. 969–974.

Bontcheva, K. and D. P. Rout (2014). "Making sense of social media streams through semantics: A survey". *Semantic Web*. 5(5): 373–403.

Boyd, D. M. and N. B. Ellison (2007). "Social network sites: Definition, history, and scholarship". *Journal of Computer-Mediated Communication*. 13(1): 210–230.

Breslin, J. G., S. Decker, A. Harth, and U. Bojars (2006). "SIOC: An approach to connect web-based communities". *Int. J. Web Based Communities*. 2(2): 133–142.

Budak, C., A. Kannan, R. Agrawal, and J. Pedersen (2014). "Inferring user interests from microblogs". In: *Technical Report, MSR-TR-2014-68*.

Çano, E. and M. Morisio (2015). "Characterization of public datasets for recommender systems". In: *2015 IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*. 249–257.

Cao, C., H. Ge, H. Lu, X. Hu, and J. Caverlee (2017). "What are you known for? Learning user topical profiles with implicit and explicit footprints". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17*. Shinjuku, Tokyo, Japan: Association for Computing Machinery. 743–752.

Cao, Y., L. Hou, J. Li, and Z. Liu (2018). "Neural collective entity linking". In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, NM, USA, August 20–26*. Ed. by E. M. Bender, L. Derczynski, and P. Isabelle. Association for Computational Linguistics. 675–686. URL: https://www.aclweb.org/anthology/C18-1057/.

Cao, Y., X. Wang, X. He, Z. Hu, and T. Chua (2019). "Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences". In: *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13–17*. 151–161.

Cavallari, S., V. W. Zheng, H. Cai, K. C. Chang, and E. Cambria (2017). "Learning community embedding with community detection and node embedding on graphs". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 6–10*. 377–386.

Ceccarelli, D., C. Lucchese, S. Orlando, R. Perego, and S. Trani (2013). "Learning relatedness measures for entity linking". In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. CIKM '13*. San Francisco, CA, USA: Association for Computing Machinery. 139–148.

Chang, Y., J. Tang, D. Yin, M. Yamada, and Y. Liu (2016). "Timeline summarization from social media with life cycle models". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. IJCAI'16*. New York, NY, USA: AAAI Press. 3698–3704.

Chen, J., J. Yu, Y. Qian, P. Li, and C. Bian (2020). "A novel recommender algorithm based on graph embedding and diffusion sampling". *Concurrency and Computation: Practice and Experience*. 32(17): e5664.

Chen, J., R. Nairn, L. Nelson, M. S. Bernstein, and E. H. Chi (2010). "Short and tweet: Experiments on recommending content from information streams". In: *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, GA, USA, April 10–15*. 1185–1194.

Chen, P. P.-S. (1976). "The entity-relationship model: Toward a unified view of data". *ACM Transactions on Database Systems*. 1: 9–36.

Chen, Y., J. Zhao, X. Hu, X. Zhang, Z. Li, and T. Chua (2013). "From interest to function: Location estimation in social media". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14–18, Bellevue, WA, USA*. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6286.

Cheng, Z., J. Caverlee, H. Barthwal, and V. Bachani (2014). "Who is the barbecue king of Texas? A geo-spatial approach to finding local experts on Twitter". In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '14*. Gold Coast, Queensland, Australia: Association for Computing Machinery. 335–344.

Cinar, Y. G., S. Zoghbi, and M. Moens (2015). "Inferring user interests on social media from text and images". In: *IEEE International Conference on Data Mining Workshop, ICDMW 2015, Atlantic City, NJ, USA, November 14–17*. 1342–1347.

Comito, C., A. Forestiero, and C. Pizzuti (2019a). "Bursty event detection in Twitter streams". *ACM Trans. Knowl. Discov. Data*. 13(4): 41:1–41:28.

Comito, C., A. Forestiero, and C. Pizzuti (2019b). "Word embedding based clustering to detect topics in social media". In: *IEEE WIC ACM International Conference on Web Intelligence. WI '19*. Thessaloniki, Greece: ACM. 192–199.

DiMaggio, P. (2003). "Culture and cognition". *Annual Review of Sociology*. 23(Nov.): 263–287.

Ding, T., W. K. Bickel, and S. Pan (2017). "Multi-view unsupervised user feature embedding for social media-based substance use prediction". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11*. 2275–2284.

Dinh, X. T. and H. Van Pham (2020). "A proposal of deep learning model for classifying user interests on social networks". In: *Proceedings of the 4th International Conference on Machine Learning and Soft Computing. ICMLSC 2020*. Haiphong City, Viet Nam: Association for Computing Machinery. 10–14.

Elmongui, H. G., R. Mansour, H. Morsy, S. Khater, A. El-Sharkasy, and R. Ibrahim (2015). "TRUPI: Twitter recommendation based on users' personal interests". In: *Computational Linguistics and Intelligent Text Processing – 16th International Conference, CICLing 2015, Cairo, Egypt, April 14–20, Proceedings, Part II*. 272–284.

Fani, H., E. Bagheri, F. Zarrinkalam, X. Zhao, and W. Du (2018). "Finding diachronic like-minded users". *Computational Intelligence*. 34(1): 124–144.

Faralli, S., G. Stilo, and P. Velardi (2015). "Recommendation of microblog users based on hierarchical interest profiles". *Social Network Analysis and Mining*. 5(1): 25.

Faralli, S., G. Stilo, and P. Velardi (2017). "Automatic acquisition of a taxonomy of microblogs users' interests". *Journal of Web Semantics*. 45: 23–40.

Fard, A. M., E. Bagheri, and K. Wang (2019). "Relationship prediction in dynamic heterogeneous information networks". In: *Advances in Information Retrieval – 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, Proceedings, Part I*. 19–34.

Farzindar, A. and W. Khreich (2015). "A survey of techniques for event detection in Twitter". *Computational Intelligence*. 31(1): 132–164.

Feng, W. and J. Wang (2013). "Retweet or not?: Personalized tweet re-ranking". In: *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4–8*. 577–586.

Feng, Y., F. Zarrinkalam, E. Bagheri, H. Fani, and F. Al-Obeidat (2018). "Entity linking of tweets based on dominant entity candidates". *Social Network Analysis and Mining*. 8(1): 46.

Ferragina, P. and U. Scaiella (2012). "Fast and accurate annotation of short texts with Wikipedia pages". *IEEE Software*. 29(1): 70–75.

Fiallos, A. and K. Jimenes (2019). "Using reddit data for multi-label text classification of Twitter users interests". In: *2019 Sixth International Conference on eDemocracy eGovernment (ICEDEG)*. 324–327.

Flati, T., D. Vannella, T. Pasini, and R. Navigli (2014). "Two is bigger (and better) than one: The Wikipedia bitaxonomy project". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22–27, Baltimore, MD, USA, Volume 1: Long Papers*. 945–955. URL: https://www.aclweb.org/anthology/P14-1089/.

Francis, W. N. and H. Kucera (1979). "Brown corpus manual". *Tech. rep.* Department of Linguistics, Brown University, Providence, RI, USA. URL: http://icame.uib.no/brown/bcm.html.

Fuchs, C. (2014). *Social Media: A Critical Introduction*. Sage.

Gao, Q., F. Abel, and G.-J. Houben (2012). "GeniUS: Generic user modeling library for the social semantic web". In: *Proceedings of the 2011 Joint International Conference on The Semantic Web. JIST'11*. Hangzhou, China: Springer-Verlag. 160–175.

Gasparetti, F. (2017). "Modeling user interests from web browsing activities". *Data Min. Knowl. Discov.* 31(2): 502–547.

Geng, X., H. Zhang, Z. Song, Y. Yang, H. Luan, and T.-S. Chua (2014). "One of a kind: User profiling by social curation". In: *Proceedings of the 22nd ACM International Conference on Multimedia. MM '14*. Orlando, FL, USA: Association for Computing Machinery. 567–576.

Ghosh, S., N. K. Sharma, F. Benevenuto, N. Ganguly, and P. K. Gummadi (2012). "Cognos: Crowdsourcing search for topic experts in microblogs". In: *The 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12–16*. 575–590.

Golbeck, J., C. Robles, and K. Turner (2011). "Predicting personality with social media". In: *CHI '11 Extended Abstracts on Human Factors in Computing Systems. CHI EA '11*. Vancouver, BC, Canada: ACM. 253–262.

Gonçalves, R., M. Horridge, R. Li, Y. Liu, M. Musen, C. Nyulas, E. Obamos, D. Shrouty, and D. Temple (2019). "Use of OWL and semantic web technologies at pinterest". In: *The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science, 11779*. Ed. by C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, and F. Gandon. Cham: Springer. 418–435.

Gong, L., L. Lin, W. Song, and H. Wang (2020). "JNET: Learning user representations via joint network embedding and topic embedding". In: *Proceedings of the 13th International Conference on Web Search and Data Mining. WSDM '20*. Houston, TX, USA: Association for Computing Machinery. 205–213.

Granger, C. W. J. (1969). "Investigating causal relations by econometric models and cross-spectral methods". *Econometrica*. 37(3): 424–438.

Graves, M., A. Constabaris, and D. Brickley (2007). "FOAF: Connecting people on the semantic web". *Cataloging & Classification Quarterly.* 43(Apr.): 191–202.

Grbovic, M., V. Radosavljevic, N. Djuric, N. Bhamidipati, and A. Nagarajan (2015). "Gender and interest targeting for sponsored post advertising at Tumblr". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15.* Sydney, NSW, Australia: Association for Computing Machinery. 1819–1828.

Grbovic, M., N. Djuric, and V. Radosavljevic (2016a). "TargetAd2016: 2nd international workshop on Ad targeting at scale". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. WSDM '16.* San Francisco, CA, USA: Association for Computing Machinery. 693–694.

Grbovic, M., V. Radosavljevic, N. Djuric, N. Bhamidipati, and A. Nagarajan (2016b). "Leveraging blogging activity on Tumblr to infer demographics and interests of users for advertising purposes". In: *Proceedings of the 6th Workshop on "Making Sense of Microposts" Co-Located with the 25th International World Wide Web Conference (WWW 2016), Montréal, Canada, April 11.* Vol. 1691. *CEUR Workshop Proceedings.* CEUR-WS.org. 2–11. URL: http://ceur-ws.org/Vol-1691/paper_15.pdf.

Große-Bölting, G., C. Nishioka, and A. Scherp (2015). "Generic process for extracting user profiles from social media using hierarchical knowledge bases". In: *Proceedings of the 9th IEEE International Conference on Semantic Computing, ICSC 2015, Anaheim, CA, USA, February 7–9.* 197–200.

Grover, A. and J. Leskovec (2016). "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17.* 855–864.

Gruber, T. (2009). "Ontology". In: *Encyclopedia of Database Systems.* Ed. by L. Liu and M. T. Özsu. Boston, MA: Springer US. 1963–1965.

Guo, G., J. Zhang, Z. Sun, and N. Yorke-Smith (2015). "LibRec: A Java library for recommender systems". In: *Posters, Demos, Late-Breaking Results and Workshop Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization (UMAP 2015), Dublin, Ireland, June 29–July 3*. URL: http://ceur-ws.org/Vol-1388/demo_paper1.pdf.

Guy, I., U. Avraham, D. Carmel, S. Ur, M. Jacovi, and I. Ronen (2013). "Mining expertise and interests from social media". In: *Proceedings of the 22nd International Conference on World Wide Web. WWW '13*. Rio de Janeiro, Brazil: Association for Computing Machinery. 515–526.

Halverson, C. (1994). *The Developing Structure of Temperament and Personality from Infancy to Adulthood*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

Han, J. and H. Lee (2016). "Characterizing the interests of social media users: Refinement of a topic model for incorporating heterogeneous media". *Inf. Sci.* 358–359: 112–128.

Hannon, J., M. Bennett, and B. Smyth (2010). "Recommending Twitter users to follow using content and collaborative filtering approaches". In: *Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26–30*. 199–206.

Hannon, J., K. McCarthy, J. Lynch, and B. Smyth (2011a). "Personalized and automatic social summarization of events in video". In: *Proceedings of the 16th International Conference on Intelligent User Interfaces. IUI '11*. Palo Alto, CA, USA: Association for Computing Machinery. 335–338.

Hannon, J., K. McCarthy, and B. Smyth (2011b). "Finding useful users on Twitter: Twittomender the followee recommender". In: *Advances in Information Retrieval – 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18–21, Proceedings*. 784–787.

Harakawa, R., D. Takehara, T. Ogawa, and M. Haseyama (2018). "Sentiment-aware personalized tweet recommendation through multimodal FFM". *Multimedia Tools and Applications*. 77(14): 18741–18759.

Harper, F. M. and J. A. Konstan (2016). "The MovieLens datasets: History and context". *TiiS*. 5(4): 19:1–19:19.

Hasibi, F., K. Balog, and S. E. Bratsberg (2016). "Exploiting entity linking in queries for entity retrieval". In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. ICTIR '16*. Newark, DE, USA: Association for Computing Machinery. 209–218.

He, J., H. Liu, Y. Zheng, S. Tang, W. He, and X. Du (2020). "Bi-Labeled LDA: Inferring interest tags for non-famous users in social network". *Data Science and Engineering*. 5: 27–47.

He, K., X. Zhang, S. Ren, and J. Sun (2015a). "Deep residual learning for image recognition". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 770–778.

He, L., Y. Jia, W. Han, and Z. Ding (2014). "Mining user interest in microblogs with a user-topic model". *China Communications*. 11(8): 131–144.

He, W., H. Liu, J. He, S. Tang, and X. Du (2015b). "Extracting interest tags for non-famous users in social network". In: *CIKM*. 861–870.

He, X., T. Chen, M.-Y. Kan, and X. Chen (2015c). "TriRank: Review-aware explainable recommendation by modeling aspects". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. CIKM '15*. Melbourne, Australia: Association for Computing Machinery. 1661–1670.

Holub, M. and M. Bieliková (2010). "Estimation of user interest in visited web page". In: *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, NC, USA, April 26–30*. 1111–1112.

Hong, L. and B. D. Davison (2010). "Empirical study of topic modeling in Twitter". In: *1st Workshop on Social Media Analytics (SOMA '10)*. 80–88.

Hong, L., A. S. Doumith, and B. D. Davison (2013). "Co-factorization machines: Modeling user interests and predicting individual decisions in Twitter". In: *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4–8*. 557–566.

Hosseini, H., T. T. Nguyen, and E. Bagheri (2018). "Implicit entity linking through ad-hoc retrieval". In: *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28–31*. 326–329.

Huang, J., M. Peng, H. Wang, J. Cao, W. Gao, and X. Zhang (2017a). "A probabilistic method for emerging topic tracking in microblog stream". *World Wide Web*. 20(2): 325–350.

Huang, S., J. Zhang, D. Schonfeld, L. Wang, and X. Hua (2017b). "Two-stage friend recommendation based on network alignment and series expansion of probabilistic topic model". *IEEE Trans. Multimedia*. 19(6): 1314–1326.

Inaba, M. and K. Takahashi (2018). "Estimating user interest from open-domain dialogue". In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. 32–40.

Jamali, M. and M. Ester (2010). "A matrix factorization technique with trust propagation for recommendation in social networks". In: *Proceedings of the Fourth ACM Conference on Recommender Systems. RecSys '10*. Barcelona, Spain: ACM. 135–142.

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell (2014). "Caffe: Convolutional architecture for fast feature embedding". In: *Proceedings of the 22nd ACM International Conference on Multimedia. MM '14*. Orlando, FL, USA: Association for Computing Machinery. 675–678.

Jin, X., W. Lei, Z. Ren, H. Chen, S. Liang, Y. Zhao, and D. Yin (2018). "Explicit state tracking with semi-supervision for neural dialogue generation". *CoRR*. abs/1808.10596. arXiv: 1808.10596. URL: http://arxiv.org/abs/1808.10596.

Joshi, D., M. Cooper, F. Chen, and Y.-Y. Chen (2015). "Building user profiles from shared photos". In: *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions. MMCommons '15*. Brisbane, Australia: ACM. 37–42.

Juan, Y., Y. Zhuang, W.-S. Chin, and C.-J. Lin (2016). "Field-aware factorization machines for CTR prediction". In: *Proceedings of the 10th ACM Conference on Recommender Systems. RecSys '16*. Boston, MA, USA: Association for Computing Machinery. 43–50.

Kabbur, S., X. Ning, and G. Karypis (2013). "FISM: Factored item similarity models for top-N recommender systems". In: *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11–14*. 659–667.

Kang, J. and H. Lee (2017). "Modeling user interest in social media using news media and Wikipedia". *Information Systems*. 65: 52–64.

Kang, J., H. Choi, and H. Lee (2019). "Deep recurrent convolutional networks for inferring user interests from social media". *J. Intell. Inf. Syst.* 52(1): 191–209.

Kapanipathi, P., F. Orlandi, A. P. Sheth, and A. Passant (2011). "Personalized filtering of the Twitter stream". In: *Proceedings of the Second Workshop on Semantic Personalized Information Management: Retrieval and Recommendation 2011, Bonn, Germany, October 24*. 6–13.

Kapanipathi, P., P. Jain, C. Venkatramani, and A. P. Sheth (2014). "User interests identification on Twitter using a hierarchical knowledge base". In: *The Semantic Web: Trends and Challenges – 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25–29, Proceedings*. 99–113.

Karatay, D. and P. Karagoz (2015). "User interest modeling in twitter with named entity recognition". *CEUR Workshop Proceedings*. 1395(Jan.): 17–20.

Kaushal, V. and M. Patwardhan (2018). "Emerging trends in personality identification using online social networks: A literature survey". *ACM Trans. Knowl. Discov. Data*. 12(2): 15:1–15:30.

Khanian Najafabadi, M., A. Mohamed, and M. Mahrin (2017). "A survey on data mining techniques in recommender systems". *Soft Computing*. 23: 627–654.

Khater, S., H. Elmongui, and D. Gracanin (2014). "Tweets you like: Personalized tweets recommendation based on dynamic users interests". In: *ASE BigData Conference, Harvard University, December 14–16*.

Kosinski, M., S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell (2015). "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines." *The American Psychologist*. 70(6): 543–56.

Kwak, H., C. Lee, H. Park, and S. B. Moon (2010). "What is Twitter, a social network or a news media?" In: *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, NC, USA, April 26–30*. 591–600.

Lazzez, O., W. Ouarda, and A. M. Alimi (2018). "DeepVisInterests: CNN-ontology prediction of users interests from social images". *CoRR*. abs/1811.10920. arXiv: 1811.10920. URL: http://arxiv.org/abs/1811.10920.

Le, Q. and T. Mikolov (2014). "Distributed representations of sentences and documents". In: *Proceedings of the 31st International Conference on International Conference on Machine Learning – Volume 32. ICML'14*. Beijing, China: JMLR.org. II-1188–II-1196. URL: http://dl.acm.org/citation.cfm?id=3044805.3045025.

Letierce, J., A. Passant, J. Breslin, and S. Decker (2010). "Understanding how Twitter is used to spread scientific messages". In: *Proceedings of Web Science Conference 2010, April 26–27, Raleigh, NC, USA*.

Li, C., C. Quan, L. Peng, Y. Qi, Y. Deng, and L. Wu (2019). "A capsule network for recommendation and explaining what you like and dislike". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'19*. Paris, France: Association for Computing Machinery. 275–284.

Li, J. and P. Zhang (2013). "Mining explainable user interests from scalable user behavior data". In: *Proceedings of the First International Conference on Information Technology and Quantitative Management*. 789–796.

Li, W., C. Eickhoff, and A. P. de Vries (2014). "Geo-spatial domain expertise in microblogs". In: *Advances in Information Retrieval*. Ed. by M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann. Cham: Springer International Publishing. 487–492.

Li, X., L. Guo, and Y. E. Zhao (2008). "Tag-based social interest discovery". In: *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21–25*. 675–684.

Liang, S., X. Zhang, Z. Ren, and E. Kanoulas (2018). "Dynamic embeddings for user profiling in Twitter". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19–23*. 1764–1773.

Lim, K. H. (2015). "Recommending tours and places-of-interest based on user interests from geo-tagged photos". In: *Proceedings of the 2015 ACM SIGMOD PhD Symposium, Melbourne, VIC, Australia, May 31–June 04*. 33–38.

Lim, K. H. and A. Datta (2012a). "Finding Twitter communities with common interests using following links of celebrities". In: *Proceedings of the 3rd International Workshop on Modeling Social Media, MSM 2012, Milwaukee, WI, USA, June 25*. 25–32.

Lim, K. H. and A. Datta (2012b). "Following the follower: Detecting communities with common interests on Twitter". In: *23rd ACM Conference on Hypertext and Social Media, HT '12, Milwaukee, WI, USA, June 25–28*. 317–318.

Lim, K. H., J. Chan, C. Leckie, and S. Karunasekera (2018). "Personalized trip recommendation for tourists based on user interests, points of interest visit durations and visit recency". *Knowledge and Information Systems*. 54(Feb.): 375–406.

Liu, Y., C. Kliman-Silver, and A. Mislove (2014). "The tweets they are a-changin: Evolution of Twitter users and behavior". In: *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, MI, USA, June 1–4*. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8043.

Liu, Y., M. Ester, B. Hu, and D. W. Cheung (2015). "Spatio-temporal topic models for check-in data". In: *2015 IEEE International Conference on Data Mining*. 889–894.

Long, R., H. Wang, Y. Chen, O. Jin, and Y. Yu (2011). "Towards effective event detection, tracking and summarization on microblog data". In: *Web-Age Information Management*. Ed. by H. Wang, S. Li, S. Oyama, X. Hu, and T. Qian. Berlin, Heidelberg: Springer. 652–663.

Lu, C., W. Lam, and Y. Zhang (2012). "Twitter user modeling and tweets recommendation based on Wikipedia concept graph". In: *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Lully, V., P. Laublet, M. Stankovic, and F. Radulovic (2018). "Exploring the synergy between knowledge graph and computer vision for personalisation systems". *Procedia Computer Science*. 137: 175–186. Proceedings of the 14th International Conference on Semantic Systems September 10–13, 2018 Vienna, Austria. URL: http://www.sciencedirect.com/science/article/pii/S1877050918316223.

Lumsden, J. (2008). *Handbook of Research on User Interface Design and Evaluation for Mobile Technology* (2 Volumes). Hershey, PA, USA. URL: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-59904-871-0.

Ly, T. H., S. T. Do, and T. T. S. Nguyen (2018). "Ontology-based recommender system for the million song dataset challenge". In: *10th International Conference on Knowledge and Systems Engineering, KSE 2018, Ho Chi Minh City, Vietnam, November 1–3*. 236–241.

Ma, Y., Y. Zeng, X. Ren, and N. Zhong (2011). "User interests modeling based on multi-source personal information fusion and semantic reasoning". In: *Active Media Technology*. Ed. by N. Zhong, V. Callaghan, A. A. Ghorbani, and B. Hu. Berlin, Heidelberg: Springer. 195–205.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mao, X.-L., Z.-Y. Ming, T.-S. Chua, S. Li, H. Yan, and X. Li (2012). "SSHLDA: A semi-supervised hierarchical topic model". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics. 800–809. URL: https://www.aclweb.org/anthology/D12-1073.

Maron, M. E., S. Curry, and P. Thompson (1986). "An inductive search system: Theory, design, and implementation". *IEEE Trans. Systems, Man, and Cybernetics*. 16(1): 21–28.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). "Birds of a feather: Homophily in social networks". *Annual Review of Sociology*: 415–444.

Mehrotra, R., S. Sanner, W. Buntine, and L. Xie (2013). "Improving LDA topic models for microblogs via tweet pooling and automatic labeling". In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '13*. Dublin, Ireland: ACM. 889–892.

Meij, E., W. Weerkamp, and M. de Rijke (2012). "Adding semantics to microblog posts". In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. WSDM '12*. Seattle, WA, USA: Association for Computing Machinery. 563–572.

Mendes, P. N., M. Jakob, A. Garcia-Silva, and C. Bizer (2011). "DBpedia spotlight: Shedding light on the web of documents". In: *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7–9*. 1–8.

Michelson, M. and S. A. Macskassy (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 2010, Toronto, Ontario, Canada, October 26th (in Conjunction with CIKM 2010)*. 73–80.

Mihalcea, R. and A. Csomai (2007). "Wikify! linking documents to encyclopedic knowledge". In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. CIKM '07*. Lisbon, Portugal: Association for Computing Machinery. 233–242.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a Meeting Held December 5–8, Lake Tahoe, Nevada, United States*. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

Miles, A., B. Matthews, M. Wilson, and D. Brickley (2005). "SKOS core: Simple knowledge organisation for the web". *International Conference on Dublin Core and Metadata Applications*: 3–10. URL: http://dcpapers.dublincore.org/pubs/article/view/79.

Milne, D. and I. H. Witten (2008). "Learning to link with Wikipedia". In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management. CIKM '08*. Napa Valley, CA, USA: Association for Computing Machinery. 509–518.

Myers, S. A. and J. Leskovec (2014). "The bursty dynamics of the Twitter information network". In: *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7–11*. 913–924.

Narducci, F., C. Musto, G. Semeraro, P. Lops, and M. de Gemmis (2013). "Leveraging encyclopedic knowledge for transparent and serendipitous user profiles". In: *User Modeling, Adaptation, and Personalization – 21th International Conference, UMAP 2013, Rome, Italy, June 10–14, Proceedings*. 350–352.

Naruchitparames, J., M. H. Gunes, and S. J. Louis (2011). "Friend recommendations in social networks using genetic algorithms and network topology". In: *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2011, New Orleans, LA, USA, 5–8 June*. 2207–2214.

Nazari, N. and M. A. Mahdavi (2019). "A survey on automatic text summarization". *Journal of AI and Data Mining*. 7(1): 121–135.

Nguyen, H., R. Richards, C.-C. Chan, and K. J. Liszka (2016). "RedTweet: Recommendation engine for reddit". *Journal of Intelligent Information Systems*. 47(2): 247–265.

Nishioka, C. and A. Scherp (2016). "Profiling vs. time vs. content: What does matter for Top-k publication recommendation based on Twitter profiles?" In: *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016, Newark, NJ, USA, June 19–23*. 171–180.

Nishioka, C., G. Große-Bölting, and A. Scherp (2015). "Influence of time on user profiling and recommending researchers in social media". In: *Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business, I-KNOW '15, Graz, Austria, October 21–23*. 9:1–9:8.

Obar, J. and S. Wildman (2015). "Social media definition and the governance challenge: An introduction to the special issue". *Telecommunications Policy*. 39(9): 745–750.

Orlandi, F., J. G. Breslin, and A. Passant (2012). "Aggregated, interoperable and multi-domain user profiles for the social web". In: *I-SEMANTICS 2012 – 8th International Conference on Semantic Systems, I-SEMANTICS '12, Graz, Austria, September 5–7*. 41–48.

Orlandi, F., P. Kapanipathi, A. P. Sheth, and A. Passant (2013). "Characterising concepts of interest leveraging linked data and the social web". In: *2013 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2013, Atlanta, GA, USA, November 17–20*. 519–526.

Ottoni, R., D. B. L. Casas, J. P. Pesce, W. Meira, C. Wilson, A. Mislove, and V. A. F. Almeida (2014). "Of pins and tweets: Investigating how users behave across image- and text-based social networks". In: *ICWSM*.

Pan, S. and T. Ding (2019). "Social media-based user embedding: A literature review". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16*. 6318–6324.

Pandey, M. and A. C. Y. Sang (2015). "Capturing the visual language of social media". In: *2015 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6.

Paulheim, H. (2016). "Knowledge graph refinement: A survey of approaches and evaluation methods". *Semantic Web*. 8(Dec.): 489–508.

Peña, P., R. D. Hoyo, J. Vea-Murguia, C. González, and S. Mayo (2013). "Collective knowledge ontology user profiling for Twitter – automatic user profiling". In: *2013 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2013, Atlanta, GA, USA, November 17–20*. 439–444.

Pennacchiotti, M., F. Silvestri, H. Vahabi, and R. Venturini (2012). "Making your interests follow you on Twitter". In: *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29–November 02*. 165–174.

Pennebaker, J., M. Francis, and R. Booth (1999). "Linguistic inquiry and word count (LIWC)". Jan.

Perozzi, B., R. Al-Rfou, and S. Skiena (2014). "DeepWalk: Online learning of social representations". In: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA – August 24–27*. 701–710.

Petkos, G., S. Papadopoulos, L. M. Aiello, R. Skraba, and Y. Kompatsiaris (2014). "A soft frequent pattern mining approach for textual topic detection". In: *WIMS*. 25:1–25:10.

Phelan, O., K. McCarthy, and B. Smyth (2009). "Using twitter to recommend real-time topical news". In: *Proceedings of the 2009 ACM Conference on Recommender Systems, RecSys 2009, New York, NY, USA, October 23–25*. 385–388.

Piao, G. and J. G. Breslin (2016a). "Analyzing aggregated semantics-enabled user modeling on Google+ and Twitter for personalized link recommendations". In: *UMAP2016*. 105–109.

Piao, G. and J. G. Breslin (2016b). "Analyzing MOOC entries of professionals on LinkedIn for user modeling and personalized MOOC recommendations". In: *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization. UMAP '16*. Halifax, Nova Scotia, Canada: Association for Computing Machinery. 291–292.

Piao, G. and J. G. Breslin (2016c). "Exploring dynamics and semantics of user interests for user modeling on twitter for link recommendations". In: *Proceedings of the 12th International Conference on Semantic Systems, SEMANTICS 2016, Leipzig, Germany, September 12–15*. 81–88.

Piao, G. and J. G. Breslin (2016d). "Interest representation, enrichment, dynamics, and propagation: A study of the synergetic effect of different user modeling dimensions for personalized recommendations on Twitter". In: *Knowledge Engineering and Knowledge Management – 20th International Conference, EKAW 2016, Bologna, Italy, November 19–23, 2016, Proceedings*. 496–510.

Piao, G. and J. G. Breslin (2016e). "User modeling on Twitter with WordNet synsets and DBpedia concepts for personalized recommendations". In: *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24–28*. 2057–2060.

Piao, G. and J. G. Breslin (2017a). "Inferring user interests for passive users on Twitter by leveraging followee biographies". In: *ECIR*. 122–133.

Piao, G. and J. G. Breslin (2017b). "Leveraging followee list memberships for inferring user interests for passive users on Twitter". In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT 2017, Prague, Czech Republic, July 4–7*. 155–164.

Piao, G. and J. G. Breslin (2018a). "Inferring user interests in microblogging social networks: A survey". *User Model. User-Adapt. Interact.* 28(3): 277–329.

Piao, G. and J. G. Breslin (2018b). "Learning to rank tweets with author-based long short-term memory networks". In: *Web Engineering*. Cham: Springer International Publishing. 288–295.

Poddar, A., E. Zangerle, and Y.-H. Yang (2018). "#nowplaying-RS: A new benchmark dataset for building context-aware music recommender systems". In: *Proceedings of the 15th Sound & Music Computing Conference*. code at URL: https://github.com/asmitapoddar/nowplaying-RS-Music-Reco-FM. Limassol, Cyprus. URL: http://mac.citi.sinica.edu.tw/~yang/pub/poddar18smc.pdf.

Ponzetto, S. P. and M. Strube (2007). "Deriving a large scale taxonomy from Wikipedia". In: *Proceedings of the 22nd National Conference on Artificial Intelligence – Volume 2. AAAI'07*. Vancouver, British Columbia, Canada: AAAI Press. 1440–1445. URL: http://dl.acm.org/citation.cfm?id=1619797.1619876.

Porteous, I., E. Bart, and M. Welling (2008). "Multi-HDP: A non parametric Bayesian model for tensor factorization". In: *AAAI*.

Purohit, H., A. Dow, O. Alonso, L. Duan, and K. Haas (2012). "User taglines: Alternative presentations of expertise and interest in social media". *2012 International Conference on Social Informatics, Lausanne, 2012*. IEEE. 236–243.

Rajani, N. F. N., K. McArdle, and J. Baldridge (2014). "Extracting topics based on authors, recipients and content in microblogs". In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '14*. Gold Coast, Queensland, Australia: ACM. 1171–1174.

Rajaraman, A. and J. D. Ullman (2011). "Data mining". In: *Mining of Massive Datasets*. Cambridge University Press. 1–17.

Rakesh, V., D. Singh, B. Vinzamuri, and C. K. Reddy (2014). "Personalized recommendation of Twitter lists using content and network information". In: *ICWSM*.

Reinanda, R., E. Meij, and M. de Rijke (2015). "Mining, ranking and recommending entity aspects". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15*. Santiago, Chile: Association for Computing Machinery. 263–272.

Ren, Z., S. Liang, E. Meij, and M. de Rijke (2013). "Personalized time-aware tweets summarization". In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '13*. Dublin, Ireland: Association for Computing Machinery. 513–522.

Ren, Z., S. Liang, P. Li, S. Wang, and M. de Rijke (2017). "Social collaborative viewpoint regression with explainable recommendations". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. WSDM '17*. Cambridge, UK: Association for Computing Machinery. 485–494.

Rohani, V. A. and O. S. Hock (2010). "On social network web sites: Definition, features, architectures and analysis tools". *Journal of Advances in Computer Research*. 1(2): 41–53.

Rohn, U. (2015). "Social media business models". In: *The International Encyclopedia of Digital Communication and Society*. American Cancer Society. 1–12.

Romero, D. M., W. Galuba, S. Asur, and B. A. Huberman (2011). "Influence and passivity in social media". In: *Machine Learning and Knowledge Discovery in Databases – European Conference, ECML PKDD 2011, Athens, Greece, September 5–9, Proceedings, Part III*. 18–33.

Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smyth (2004). "The author-topic model for authors and documents". In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. UAI '04*. Banff, Canada: AUAI Press. 487–494. URL: http://dl.acm.org/citation.cfm?id=1036843.1036902.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li (2015). "ImageNet large scale visual recognition challenge". *International Journal of Computer Vision*. 115(3): 211–252.

Safari, R. M., A. M. Rahmani, and S. H. Alizadeh (2019). "User behavior mining on social media: A systematic literature review". *Multimedia Tools and Applications*. 78: 33747–33804.

Sang, J., D. Lu, and C. Xu (2015). "A probabilistic framework for temporal user modeling on microblogs". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. CIKM '15*. Melbourne, Australia: ACM. 961–970.

Sasaki, K., T. Yoshikawa, and T. Furuhashi (2014). "Online topic model for Twitter considering dynamics of user interests and topic trends". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics. 1977–1985. URL: https://www.aclweb.org/anthology/D14-1212.

Seghouani, N. B., C. N. Jipmo, and G. Quercini (2018). "Determining the interests of social media users: Two approaches". *Information Retrieval Journal*: 1–30.

Shadbolt, N., T. Berners-Lee, and W. Hall (2006). "The semantic web revisited". *IEEE Intelligent Systems*. 21(3): 96–101.

Shen, W., J. Wang, P. Luo, and M. Wang (2013). "Linking named entities in tweets with knowledge base via user interest modeling". In: *KDD*. 68–76.

Shi, L., L. Liu, Y. Wu, L. Jiang, and J. Hardy (2017). "Event detection and user interest discovering in social media data streams". *IEEE Access*. 5: 20953–20964.

Shu, K., S. Wang, J. Tang, R. Zafarani, and H. Liu (2016). "User identity linkage across online social networks: A review". *SIGKDD Explorations*. 18(2): 5–17.

Souri, A., S. Hosseinpour, and A. M. Rahmani (2018). "Personality classification based on profiles of social networks' users and the five-factor model of personality". *Human-Centric Computing and Information Sciences*. 8(1): 24.

Spasojevic, N., J. Yan, A. Rao, and P. Bhattacharyya (2014). "LASTA: Large scale topic assignment on multiple social networks". In: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA – August 24–27*. 1809–1818.

Sriram, B., D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas (2010). "Short text classification in twitter to improve information filtering". In: *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19–23*. 841–842.

Steyvers, M., P. Smyth, M. Rosen-Zvi, and T. L. Griffiths (2004). "Probabilistic author-topic models for information discovery". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, August 22–25*. 306–315.

Suominen, H. (2009). "Performance evaluation measures for text mining". In: *Handbook of Research on Text and Web Mining Technologies*. Hershey, PA, USA: IGI Global. 724–747.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2015). "Going deeper with convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2016). "Rethinking the inception architecture for computer vision". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016*. 2818–2826.

Tang, X. and C. C. Yang (2012). "TUT: A statistical model for detecting trends, topics and user interests in social media". In: *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29–November 02*. Ed. by X. Chen, G. Lebanon, H. Wang, and M. J. Zaki. ACM. 972–981.

Tommaso, G. D., S. Faralli, G. Stilo, and P. Velardi (2018). "Wiki-MID: A very large multi-domain interests dataset of Twitter users with mappings to Wikipedia". In: *17th International Semantic Web Conference*. 36–52.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). "Feature-rich part-of-speech tagging with a cyclic dependency network". In: *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27–June 1*. URL: https://www.aclweb.org/anthology/N03-1033/.

Trikha, A. K., F. Zarrinkalam, and E. Bagheri (2018). "Topic-association mining for user interest detection". In: *ECIR*. 665–671.

Varga, A., A. E. C. Basave, M. Rowe, F. Ciravegna, and Y. He (2014). "Linked knowledge sources for topic classification of microposts: A semantic graph-based approach". *J. Web Sem.* 26: 36–57.

Wagner, C., V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier (2012). "It's not in their Tweets: Modeling topical expertise of Twitter users". In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*. 91–100.

Wang, H., F. Zhang, M. Zhao, W. Li, X. Xie, and M. Guo (2019a). "Multi-task feature learning for knowledge graph enhanced recommendation". In: *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13–17, 2019*. 2000–2010.

Wang, H., F. Zhang, M. Hou, X. Xie, M. Guo, and Q. Liu (2018a). "SHINE: Signed heterogeneous information network embedding for sentiment link prediction". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18*. Marina Del Rey, CA, USA: Association for Computing Machinery. 592–600.

Wang, J., W. X. Zhao, Y. He, and X. Li (2014). "Infer user interests via link structure regularization". *ACM TIST*. 5(2): 23:1–23:22.

Wang, N., H. Wang, Y. Jia, and Y. Yin (2018b). "Explainable recommendation via multi-task learning in opinionated text data". In: *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '18*. Ann Arbor, MI, USA: Association for Computing Machinery. 165–174.

Wang, T., H. Liu, J. He, and X. Du (2013). "Mining user interests from information sharing behaviors in social media". In: *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14–17, Proceedings, Part II*. 85–98.

Wang, X., L. Zhang, X. Li, and W. Ma (2008). "Annotating images by mining image search results". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 30(11): 1919–1932.

Wang, X.-J., M. Yu, L. Zhang, R. Cai, and W.-Y. Ma (2009). "Argo: Intelligent advertising by mining a user's interest from his photo collections". In: *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising. ADKDD '09*. Paris, France: ACM. 18–26.

Wang, X., X. He, F. Feng, L. Nie, and T.-S. Chua (2018c). "TEM: Tree-Enhanced embedding model for explainable recommendation". In: *Proceedings of the 2018 World Wide Web Conference. WWW '18*. Lyon, France: International World Wide Web Conferences Steering Committee. 1543–1552.

Wang, X., X. He, Y. Cao, M. Liu, and T. Chua (2019b). "KGAT: Knowledge graph attention network for recommendation". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4–8*. 950–958.

Wei, W., G. Cong, C. Miao, F. Zhu, and G. Li (2016). "Learning to find topic experts in Twitter via different relations". *IEEE Trans. Knowl. Data Eng.* 28(7): 1764–1778.

Weng, J., E. Lim, J. Jiang, and Q. He (2010). "TwitterRank: Finding topic-sensitive influential twitterers". In: *WSDM*. 261–270.

Wieczorek, S., D. Filipiak, and A. Filipowska (2018). "Semantic image-based profiling of users' interests with neural networks". In: *Emerging Topics in Semantic Technologies – ISWC 2018 Satellite Events [Best Papers from 13 of the Workshops Co-Located with the ISWC 2018 Conference]*. 179–190.

Workie, A., R. S. Rajendran, and Y. Chung (2020). "Digital video summarization techniques: A survey". *International Journal of Engineering and Technology*. 09(Jan.): 5.

Xu, J. and T. Lu (2015). "Inferring user interests on Tumblr". In: *Social Computing, Behavioral-Cultural Modeling, and Prediction – 8th International Conference, SBP 2015, Washington, DC, USA, March 31–April 3. Proceedings*. Ed. by N. Agarwal, K. Xu, and N. Osgood. Vol. 9021. *Lecture Notes in Computer Science*. Springer. 458–463.

Xu, Y. (2018). "User expertise modelling using social network data". Thesis. Trinity College Dublin. School of Computer Science & Statistics.

Xu, Z., R. Lu, L. Xiang, and Q. Yang (2011). "Discovering user interest on Twitter with a modified author-topic model". In: *Proceedings of the 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011, Campus Scientifique de la Doua, Lyon, France, August 22–27*. 422–429.

Xu, Z., Y. Zhang, Y. Wu, and Q. Yang (2012). "Modeling user posting behavior on social media". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12*. Portland, OR, USA: Association for Computing Machinery. 545–554.

Xu, Y., D. Zhou, and S. Lawless (2016). "Inferring your expertise from Twitter: Integrating sentiment and topic relatedness". In: *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 121–128.

Xu, Y., D. Zhou, and S. Lawless (2017). "User expertise inference on Twitter: Learning from multiple types of user data". In: *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. UMAP '17*. Bratislava, Slovakia: Association for Computing Machinery. 395–396.

Yan, X., J. Guo, Y. Lan, and X. Cheng (2013). "A biterm topic model for short texts". In: *Proceedings of the 22nd International Conference on World Wide Web. WWW '13*. Rio de Janeiro, Brazil: ACM. 1445–1456.

Yan, X., J. Guo, Y. Lan, J. Xu, and X. Cheng (2015). "A probabilistic model for bursty topic discovery in microblogs". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25–30, Austin, TX, USA*. 353–359. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9816.

Yang, L., T. Sun, M. Zhang, and Q. Mei (2012). "We know what @you #tag: Does the dual role affect hashtag adoption?" In: *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16–20*. 261–270.

Yao, J., Y. Zhang, I. Tsang, and J. Sun (2017). "Discovering user interests from social images". In: *MultiMedia Modeling*. Ed. by L. Amsaleg, G. Þ. Guðmundsson, C. Gurrin, B. Þ. Jónsson, and S. Satoh. Cham: Springer International Publishing. 160–172.

Yin, H., B. Cui, L. Chen, Z. Hu, and X. Zhou (2015). "Dynamic user modeling in social media systems". *ACM Transactions on Information Systems*. 33(3): 10:1–10:44.

You, Q., S. Bhatia, and J. Luo (2016). "A picture tells a thousand words—About you! User interest profiling from user generated visual content". *Signal Processing*. 124: 45–53. Big Data Meets Multimedia Analytics. URL: http://www.sciencedirect.com/science/article/pii/S0165168415003758.

Youyou, W., M. Kosinski, and D. Stillwell (2015). "Computer-based personality judgments are more accurate than those made by humans". *Proceedings of the National Academy of Sciences*. 112(4): 1036–1040.

Yuan, G., P. K. Murukannaiah, Z. Zhang, and M. P. Singh (2014). "Exploiting sentiment homophily for link prediction". In: *Proceedings of the 8th ACM Conference on Recommender Systems. RecSys '14*. Foster City, Silicon Valley, CA, USA: Association for Computing Machinery. 17–24.

Zarrinkalam, F. and E. Bagheri (2016). "Event identification in social networks". *CoRR*. abs/1606.08521. arXiv: 1606.08521. URL: http://arxiv.org/abs/1606.08521.

Zarrinkalam, F., H. Fani, E. Bagheri, M. Kahani, and W. Du (2015). "Semantics-enabled user interest detection from Twitter". In: *International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6–9, 2015 – Volume I*. 469–476.

Zarrinkalam, F., H. Fani, E. Bagheri, and M. Kahani (2016). "Inferring implicit topical interests on Twitter". In: *Advances in Information Retrieval – 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23. Proceedings*. 479–491. DOI: 10.1007/978-3-319-30671-1_35.

Zarrinkalam, F., H. Fani, E. Bagheri, and M. Kahani (2017). "Predicting users' future interests on Twitter". In: *Advances in Information Retrieval – 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8–13, Proceedings*. 464–476. DOI: 10.1007/978-3-319-56608-5_36.

Zarrinkalam, F., M. Kahani, and E. Bagheri (2018). "Mining user interests over active topics on social networks". *Inf. Process. Manage.* 54(2): 339–357.

Zarrinkalam, F., H. Fani, and E. Bagheri (2019a). "Social user interest mining: Methods and applications". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19*. Anchorage, AK, USA: Association for Computing Machinery. 3235–3236. ISBN: 9781450362016. DOI: 10.1145/3292500.3332279.

Zarrinkalam, F., M. Kahani, and E. Bagheri (2019b). "User interest prediction over future unobserved topics on social networks". *Inf. Retr. Journal*. 22(1–2): 93–128.

Zhang, Y. and X. Chen (2020). "Explainable recommendation: A survey and new perspectives". *Foundations and Trends® in Information Retrieval*. 14(1): 1–101.

Zhang, Y., Y. Wu, and Q. Yang (2012). "Community discovery in Twitter based on user interests". *Journal of Computational Information Systems*. 8(Mar.).

Zhang, Y., G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma (2014). "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis". In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '14*. Gold Coast, Queensland, Australia: Association for Computing Machinery. 83–92. ISBN: 9781450322577. DOI: 10.1145/2600428.2609579.

Zhang, D., J. Yin, X. Zhu, and C. Zhang (2017). "User profile preserving social network embedding". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence. IJCAI'17*. Melbourne, Australia: AAAI Press. 3378–3384. ISBN: 978-0-9992411-0-3. URL: http://dl.acm.org/citation.cfm?id=3172077.3172361.

Zhao, W. X., J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li (2011). "Comparing Twitter and traditional media using topic models". In: *ECIR*. 338–349.

Zhao, Z., Z. Cheng, L. Hong, and E. H. Chi (2015). "Improving user topic interest profiles by behavior factorization". In: *Proceedings of the 24th International Conference on World Wide Web. WWW '15*. Florence, Italy: International World Wide Web Conferences Steering Committee. 1406–1416. ISBN: 9781450334693. DOI: 10.1145/2736277.2741656.

Zheng, N., S. Song, and H. Bao (2015). "A temporal-topic model for friend recommendations in Chinese microblogging systems". *IEEE Trans. Systems, Man, and Cybernetics: Systems*. 45(9): 1245–1253.

Zheng, X., W. Zheng, Y. Yang, W. Guo, and V. Chang (2019). "Clustering based interest prediction in social networks". *Multimedia Tools and Applications*. 78(23): 32755–32774.