There is a fine Line between Personalization and Surveillance: Semantic User Interest Tracing via Entity-level Analytics

Amit Kumar
Université de Caen Normandie
Caen, France
amit.kumar@unicaen.fr

Marc Spaniol
Université de Caen Normandie
Caen, France
marc.spaniol@unicaen.fr

ABSTRACT

User interest tracing is a common practice in many Web use-cases including, but not limited to, search, recommendation or intelligent assistants. The overall aim is to provide the user a personalized "Web experience" by aggregating and exploiting a plenitude of user data derived from collected logs, accessed contents, and/or mined community context. As such, fairly basic features such as terms and graph structures can be utilized in order to model a user's interest. While there are clearly positive aspects in the before mentioned application scenarios, the user's privacy is highly at risk. In order to highlight inherent privacy risks, this paper studies Semantic User Interest Tracing (SUIT in short) by investigating a user's publishing/editing behavior of Web contents. In contrast to existing approaches, SUIT solely exploits the (semantic) concepts [categories] inherent in documents derived via entity-level analytics. By doing so, we raise Web contents to the entity-level. Thus, we are able to abstract the user interest from plain text strings to "things". In particular, we utilize the inherited structural relationships present among the concepts derived from a knowledge graph in order to identify the user associated with a specific Web content. Our extensive experiments on Wikipedia show that our approach outperforms state of the art approaches in tracing and predicting user behavior in a single language. In addition, we also demonstrate the viability of our semantic (language-agnostic) approach in multilingual experiments. As such, SUIT is capable of revealing the user's identity, which demonstrates the fine line between personalization and surveillance, raising questions regarding ethical considerations at the same time.

CCS CONCEPTS

• **Information systems → World Wide Web**; **Web mining**; **Personalization**; *Information systems applications*; • **Security and privacy** → *Human and societal aspects of security and privacy*.

KEYWORDS

User Interest Tracing, Entity-level Analytics, Graph Neural Networks, Privacy, Ethics.

ACM Reference Format:
Amit Kumar and Marc Spaniol. 2022. There is a fine Line between Personalization and Surveillance: Semantic User Interest Tracing via Entity-level Analytics. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3501247.3531592

1 INTRODUCTION

1.1 Motivation and Problem

Even after three decades of World Wide Web, one can still realize the tremendous amount of proliferation of Web data being generated and, subsequently, being accessible to Web users. In particular, the Web 2.0 and its social networking services such as Twitter, Facebook, online discussion forums, or Wikipedia have created the so-called "prosumer" [37]: a (Web) user that actively produces and consumes. As a result, millions of new Web contents are being generated on a daily basis. However, not each and every piece of information is equally relevant to a user. In general, an average user is interested in a certain set of Web documents, only. This observation is being exploited by various personalization services and recommender systems, such as Google news feeds[1] or Amazon recommendations[2]. However, there is only a fine line between personalization and surveillance.

In this paper, we postulate that a user can be characterized by the concepts s/he is interested in or, rephrasing it more drastically: "Tell Me what You like and I will tell You Who You are". To be concise, we claim that the (semantic) concepts [categories] inherent in documents published and/or modified by a user can be utilized in order to allow the tracing of his/her interests. To this end, we raise user tracing to the entity-level and offer a novel, purely semantic, and language-agnostic approach. By doing so, our approach is capable of effectively [through (semantic) concepts] and efficiently [via a relatively small amount of training data] identifying user interest traces. While we consider the personalization of a user's Web experience (in general) as a positive thing, we also want to raise awareness of the inherent privacy problems.

1.2 Approach and Contribution

Nowadays, identification and tracing of user's interests from social media platforms texts has become a significant research topic [13]. However, it is incredibly challenging to capture user interests without categorical information. Moreover, the identification of an author for a given document has several applications in various domains such as information retrieval, bibliometrics, and plagiarism

WebSci '22, June 26–29, 2022, Barcelona, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531592

[1] https://news.google.com/foryou

[2] https://www.amazon.com/gp/help/customer/display.html?nodeId=GE4KRSZ4KAZZB4BV

detection [31–33]. The objective shared by the mentioned research topics is to identify whether a Web content can be associated with the publishing/editing behavior of a specific user. In this paper, we, therefore, introduce Semantic User Interest Tracing (SUIT in short), which aims at exploiting the (semantic) concepts [categories] inherent in documents in order to identify the user "behind" the content. To this end, SUIT identifies the concepts associated with a user in order to trace and - ultimately - reveal the publishing pattern. Further, SUIT utilizes the inherent structure and relationships among the (semantic) concepts derived from a knowledge graph in order to identify and reveal the respective/individual user interests.

Therefore, we investigate the concerned concepts based on the editing behavior of newly generated or published Web documents from Web users. We employ a novel graph convolutional network architecture (GCN) to capture the inherent characteristics among the concepts extracted from a knowledge base (KB) in order to distill the user publishing/editing patterns. We perform our experiments in multiple languages, *i.e.*, English, German, and French. In particular, we utilize the Wikipedia articles published/edited by the Wikipedia user community. Extensive experiments on a multi-language dataset demonstrate the viability of our proposed approach. Furthermore, enhanced performance over all the mentioned languages confirms our hypothesis that our purely semantic approach can be accommodated for any of the languages.

In summary, the salient contributions of this paper are:

- a language agnostic semantic user interest tracing and prediction model;
- the creation of a user interest tracing dataset based on the publishing behavior of Wikipedia editors;
- the adaptation of a GCN architecture in order to identify the structural patterns present among the (semantic) concepts linked with the different users;
- a comprehensive experimental study in multiple languages (English, German, and French) on semantic user interest tracing demonstrating the superior quality of our approach over state-of-the-art implementations and revealing a potential privacy intrusion;
- a critical reflection on the potential and risks of semantically-driven user interest tracing.

The structure of the rest of the paper is organized as follows: We present an overview of related work in Section 2. Section 3 presents the general conceptual approach for the underlying computational models. Subsequently, the sub-user representation is discussed in Section 4. After that, we give a detailed explanation of the implemented models in Section 5. Section 6 describes the experimental setup, results, and findings. A critical reflection on ethical considerations is given in Section 7. Finally, Section 8 concludes the studies of SUIT and outlines further research directions.

2 RELATED WORK & BACKGROUND

This section provides an overview of the related studies which are relevant to our research. We aggregate it into several sub-groups.

Ontology Based Models

Wide varieties of tasks related to information retrieval and natural language processing have accomplished improvement in the performance by exploiting ontologies [3, 24]. Elberrichi et al. exploit WordNet concepts for solving the task of content categorization [9]. SEMANNOREX provides semantic search over a given corpus via an underlying ontology [21]. In [29], the authors attempt to predict the next visit of a patient by exploiting the ontology and clinical history of a patient. Human character has been designed to predict its behavior in a given situation based on an ontology [8]. LOVBench [19] analyzes the user behavior based on an ontology search. However, none of the systems address the issue of user interest tracing or article authorship.

Graph Neural Networks

Lately, the concept of Graph Neural Networks (GNNs) has been widely accepted by many researchers because it has demonstrated to be beneficial across several tasks in multiple domains [43]. Some of the recent studies include traffic flow prediction [40], social recommendation system [11], fraud detection [23], and three dimensional object detection [34]. In their revolutionary work [18], the authors adapt a well variation of convolutional neural network (CNN) which works precisely on the graphs called graph convolutional networks (GCN). These networks accomplished very promising results on many benchmark datasets related to a graph. Recently, Veličković et al. introduce the development in the GNN architectures by employing the self-attention layers to avoid the shortcomings of graph convolution [38]. Graph Attention Networks (GATs) provides different weights to different nodes in the neighborhood in order to get the current node representation. In [12], the authors utilize GNNs to solve the out-of-knowledge-base (OOKB) entity problem in the general KBC setting. Kumar et al. aim to identify the most suitable entity-type for an entity based on GCN [20]. The authors of [46] introduce a system which utilizes GCN to learn the user representation and also propose the profitable relationship between the fraudster task and the recommendation task. Wang et al. learn the user personality representation using GCN [41]. Their graph is based on user document, document word and word co-occurrence relations. MeatTP utilizes GCN to predict the topic of user interest [47]. It exploits the user posting content and the interest of user social friends. In our current work, we adapt the GCN to solve the task of user interest tracing by utilizing the concepts [categories] information of Web contents.

User Profile Generation

Several user profiles have been generated by analyzing the visited Web contents of the users [10, 30]. Nevertheless, the advancement of social media platforms has transferred the interest of profile generation systems towards users' interactions on these platforms. These systems exploit either topic modeling [42] or bag-of-words [6] approaches to create the user profiles. The task of user interest representation has been addressed across various social media platforms by employing internal as well as external data sources (such as Wikipedia, mainstream news) [44]. Ottoni et al. investigate the user interests across several social media platforms [28]. They study the user interests based on Twitter and Pinterest. The authors in [17] identified user interests by exploring the Wikipedia category graph for the Twitter dataset. In [14], the authors project the social media contents into the corresponding categories of a news corpus. They estimate user interests by considering both the features of social media content and news categories. Kang et al. combine CNN and bidirectional gated recurrent unit (biGRU) to predict user interests on social media platforms [16].

Doppelgänger Detection
The authors in [15] propose the combination of different time specific features for Doppelgänger detection and called it "timeprints". They perform several experiments based on stylometric features (e.g., syntactic, lexical, domain-specific features, etc.) in combination with timeprints and show the significance of the time specific features. In [22], the authors make use of friendship networks to identify the users. Their experiments show that the contribution of 1-hop neighbors is much higher than the other similarities. Contrary to the approaches based on metadata, stylometric approaches target purely based on the user generated content. Abbasi and Chen [1] proposed a rich set of stylometric features (e.g., structural, idiosyncratic facets, etc.) and developed the "writeprints" technique for the identification of user identities. The current development of online communication augmented the rich set of stylometric features by including domain-specific aspects, such as the use of emoticons [7], favorable votes [27], and word sentiment [7]. Doppelgänger Finder [2] extracts the stylometric features and generates a score based on the similarity of writing style for each author pair.

Authorship Attribution
Authorship attribution is a well-known task and accomplishes encouraging performance in larger texts, such as blog posts and book chapters [36]. The authors in [45] proposed a semantic model based on word dependency relations and non-subject stylistic words to identify the author for unstructured texts. Villar-Rodrigueza et al. [39] proposed a feature selection technique on the linguistic features extracted for short messages and developed models in combination with supervised learning algorithms. Sousa Silva et al. [35] introduced a set of personalized and idiosyncratic stylistic markers such as emoticons, punctuations, abbreviations, etc. to train the SVM model for authorship attribution.

In summary, the above-mentioned approaches either depend on the metadata information of users or linguistics features extracted from the edited documents. So, these approaches are language and domain-dependent. On the contrary, our approach solely utilizes the concepts [categories] of the documents and exploits the inherent semantics among these concepts [categories] in order to derive the user interest patterns. Thus, it is independent of any language or domain. Therefore, our approach addresses a similar problem but is not directly comparable to the previously mentioned approaches.

3 CONCEPTUAL APPROACH

For the user's interests tracing task, we propose a methodology that receives a set of documents for different users as input and predicts those documents' potential candidate users/authors as output based on identified user publishing/editing patterns. As such, we provide a prediction module to determine whether a document is likely to be edited by a specific user entirely based on concepts [categories] of the document. Let u be the set of users (cf. Eq. 1) and d be the set of documents (cf. Eq. 2). u_{i_d} represents the set of documents associated with user u_i as shown in Equation 3.

$$u = \{u_1, u_2, \ldots, u_I\} \tag{1}$$

$$d = \{d_1, d_2, \ldots, d_P\} \tag{2}$$

$$u_{i_d} = \{d_{i_1}, d_{i_2}, \ldots, d_{i_N}\}, i \in [1, I] \tag{3}$$

With the emergence of Linked Open Data (LOD), many documents have already been interlinked/classified via an underlying ontology (e.g., Wikipedia category structure). In contrast to the previously mentioned approaches (cf. Sec. 2), we exploit such an underlying ontology which has been extracted from the YAGO KB [25]. In particular, we utilize the WordNet category system underlying YAGO. Each unit of the category system is termed a "concept". It is worth to mention here that the concepts within the WordNet category system form a directed acyclic graph (DAG). It entails that a concept might be associated with more than one super concept. The semantic user interest tracing task consists of two building blocks: sub-user representation (cf. Sec. 4) and user interest tracing model (cf. Sec. 5). The sub-user representation is needed in order to serve as a ground truth for our experiments later on in order to connect documents with users. Since an individual document's semantic representation is comparatively sparse (around 5-10 concepts compared with around 70,000 concepts of the entire ontology), we construct an aggregated sub-user representation. As a result, we obtain a sub-user representation graph, which is a DAG that forms the backbone of our GCN based approaches. This representation will subsequently serve as input for the different user tracing models. Let U_i denote the set of sub-users corresponding to user u_i and $u^j_{i_d}$ representing the set of documents associated with sub-user u^j_i as shown in Equations 4 and 5, respectively.

$$U_i = \{u^1_i, u^2_i, \ldots, u^J_i\}, i \in [1, I] \tag{4}$$

$$u^j_{i_d} = \{d^j_{i_1}, d^j_{i_2}, \ldots, d^j_{i_Q}\}, i \in [1, I], j \in [1, J] \tag{5}$$

Further, we define a function ϕ, which predicts if u^j_i is Doppelgänger of user u_i or not. A Doppelgänger represents a double or an apparition of an alive person in fiction or folklore. Equation 6 describes the formulation.

$$\phi(u^j_i, u_i) = \begin{cases} 1, & \text{if } u^j_i \text{ is Doppelgänger w.r.t. } u_i \\ 0, & \text{if } u^j_i \text{ is not Doppelgänger w.r.t. } u_i \end{cases} \tag{6}$$

$$\forall u^j_i \in U_i, i \in [1, I], j \in [1, J]$$

Given a document d_p, predict the potential candidate users/authors $\sigma(d_p)$, *i.e.*, the set of users/authors that are likely to publish/edit the corresponding document d_p. Formally, it can be defined as:

$$\sigma(d_p) = \{u_i, u_i \in u \mid \phi(u^j_i, u_i) = 1, d_p \in u^j_{i_d}\} \tag{7}$$

4 SUB-USER REPRESENTATION

In this section, we present the methodology for sub-user representation. Figure 1 depicts the conceptual approach of the sub-user representation by employing five generic steps, as follows:

1) Document Collection
In the first step, we extract the documents published/edited from

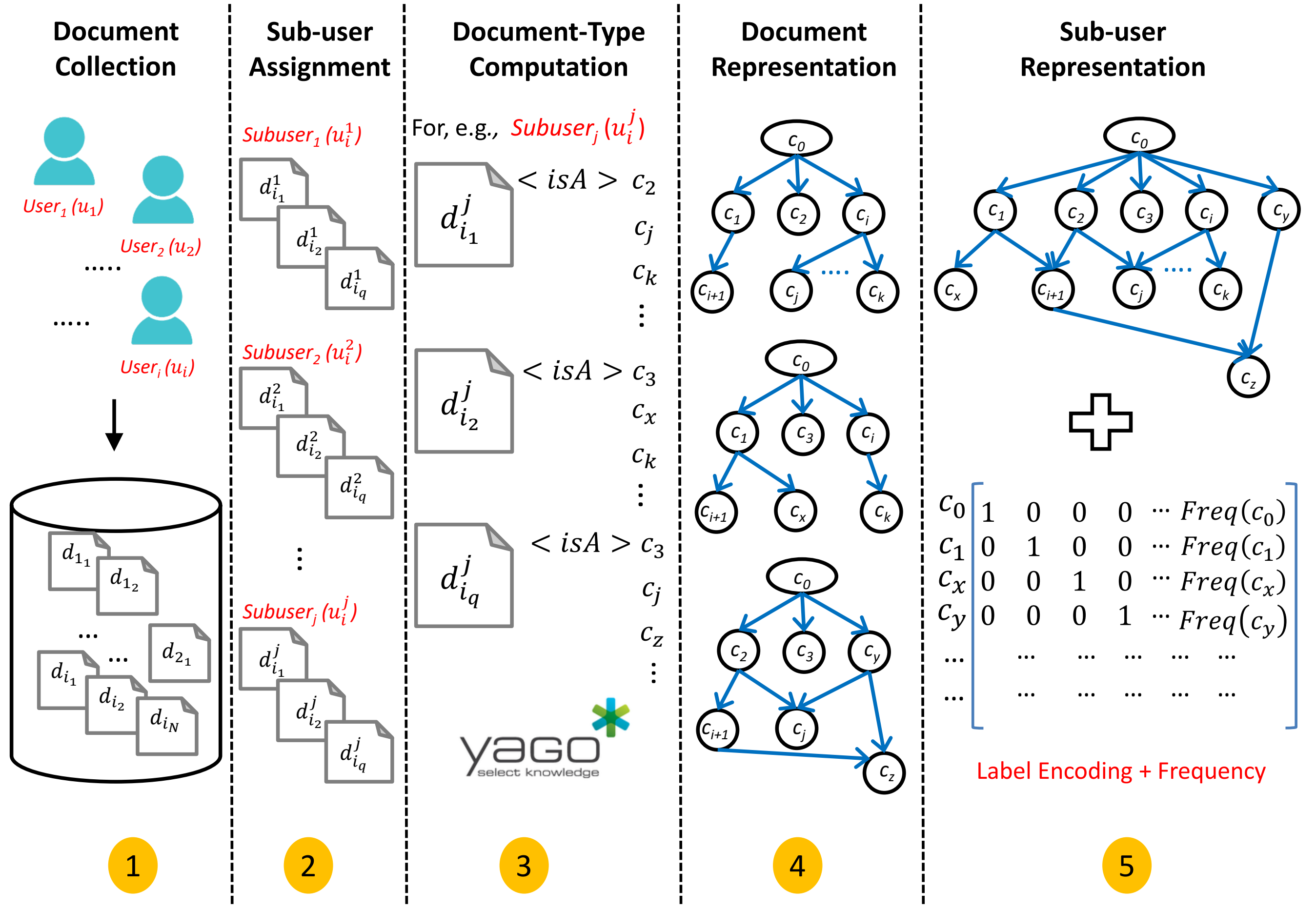


Figure 1: Conceptual Approach of Sub-user Representation

the Wikipedia revision history and create a separate list of documents associated with each user (① of Fig. 1).

2) Sub-user Assignment
The second step involves the partition of a user into J sub-users (cf. ② in Fig. 1). It means that the documents associated with a user are randomly split into J equal parts, and each part is assigned to a sub-user associated with the corresponding user. Thus, each sub-user is randomly assigned an equal number of distinguished documents from the retrieved documents associated with the respective users.

3) Document-Type Computation
For each document $d_{i_q}^j$ associated with sub-user u_i^j, we compute its associate concepts (from the underlying WordNet ontology) using the `rdf:type` relation in the YAGO KB. These concepts are called "direct concepts" (③ of Fig. 1).

4) Document Representation
In order to get the document representation, we first derive all the transitive concepts by using the `subClassOf` relation associated with the concepts computed in the previous step via the KB. Then, with the help of inherent hierarchical relationships present among the concepts (both, direct and transitive), we construct a graph. This graph consists of concepts as nodes and relationships among these concepts as edges. Again, this graph is directed and acyclic in nature. A separate graph is constructed for each document as shown in ④ of Figure 1.

5) Sub-user Representation
In the fifth and last step, all the derived concepts in the previous steps (*i.e.*, concepts as well as their transitives) along with their semantic relationships are combined. They form a larger directed acyclic graph. In addition to the graph, we also define concept label encoding (cf. Sec. 5.2) and the frequency of the respective concept (cf. Sec. 5.3) as illustrated in ⑤ of Fig. 1. We construct an individual graph for each sub-user and call it sub-user representation graph. This way, each sub-user is represented by a DAG.

Once the sub-user representation graph is constructed, it is provided as input to the next building block, *i.e.*, to the user interest tracing model. The model then aims at identifying the user publishing/editing patterns based on the sub-user representation graph. The Web documents utilized in the construction of the sub-user representation are regarded as the documents published/edited by the user who predicts the sub-user as a Doppelgänger.

5 USER INTEREST TRACING MODELS

In this section, we introduce and explain various user interest tracing models. We develop two machine learning models based on random forests. In order to learn the semantic characteristics, we also develop two graph convolutional networks based models.

5.1 Random Forest based Models

In the first step, we employ a random forest as a learner to predict the user publishing/editing patterns. Random forest (*RF*) is a classifier based on ensemble learning techniques, which utilizes a collection of several decision trees to reduce the training error [4]. The model attempts to enhance generalization by employing bootstrap aggregation over the training data and random subspace over the features.

Since a Web document can be edited by multiple users, we employ a multi-label classification technique. To this end, we convert the user interest tracing task into a set of sub-tasks and exploit the one-against-all scheme to solve the sub-tasks. Following the scheme, we train an individual classifier for each of the $|u|$ users. We follow the bag-of-words technique for the feature set construction and call it "bag-of-concepts". The size of a feature vector for a test instance depends on all the possible concepts in that particular approach. In order to encode a sub-user representation graph, we insert a "1" in the feature vector at the corresponding locations of concepts present in the graph, and the rest of the entries are set to "0". Once the feature vectors for all the sub-users are encoded, a separate *RF* classifier is trained for each of the individual users using the one-against-all scheme. Here each classifier decides if the test instance sub-user is a Doppelgänger or not. In the end, all the classifiers collectively decide the prediction for the tested sub-user.

Direct Concepts (σ_{Dir})

In our first approach, we attempt to solve the task by considering the directly connected concepts for all the Web documents associated with a sub-user as the features (cf. ③ in Sec. 4). Leaf concepts are called "direct concepts". The direct concepts that correspond to each document for train-sub-users are defined as the possible concepts in this approach. The size of a feature vector is the count of all the possible direct concepts. For example, $c_2, c_{i+1}, c_j, c_k, c_3, c_x, c_z$ are the direct concepts of the example in Figure 1 ③ and ④. We derive the feature vector for an instance using direct concepts and as discussed in the above section. The key idea behind this approach is that a Web user can be individualized by the concepts they are interested in. So, documents related to those concepts are more likely to be published/edited by that user.

Transitive Concepts (σ_{Trans})

The previously described *RF* based model (σ_{Dir}) considers only the direct concepts for the Web documents associated with a sub-user. It does not consider the other concepts which are related to those direct concepts and does not appear in the list of direct concepts for the respective sub-users. For instance, if some user is interested in topics related to the (sub-)concept `VICE_PRESIDENT` then it is most likely that the user will also be interested in topics related to the more generic concept `PRESIDENT`.

In order to address the scenario described before, we compute all the transitive closures associated with the direct concepts for all the documents corresponding to some sub-user and utilize them as the feature set. All the transitive closures concepts and their respective direct concepts for all the sub-users in the training set are defined as all the possible concepts. Again, we derive the feature vector for a sub-user using all the direct as well as transitive closure concepts associated with all the documents corresponding to that sub-user.

The fundamental limitation of the random forest based models is that they are not efficient enough to learn the inherent semantic patterns present among the concepts through the hierarchical relationships. Furthermore, the representative feature vector for the sub-user is not very informative in nature due to its sparsity.

5.2 Graph Convolutional Networks Models

In order to overcome the constraints of *RF* based models, as pointed out in the previous section, we introduce graph convolutional networks (GCN) as models for the user interest tracing task. At first, we adapt the GCN architecture for user publishing/editing patterns prediction and propose the underlying framework. The basic prediction model is represented by σ_{GCN}. We also propose an increment over the basic GCN framework represented by σ_{SUIT} in the following section by incorporating frequency information. We develop a single GCN classifier for each GCN-based model, which predicts the potential users/authors for a given document. The GCN models aim at learning the patterns by employing the sub-user representation graph and its associated concepts. More specifically, these models exploit the inherent semantics relationship among the concepts derived from the YAGO KB.

Graph Convolutional Networks

A GCN is a multi-layered neural network architecture that precisely operates on a graph designed dataset and generates an embedding vector associated with each node of the graph [18]. These embedding vectors are based on the attributes of the direct neighbors of the nodes. One GCN layer provides information about only the direct neighborhood of the nodes. Information about broader proximity can be integrated by the stacking of multiple GCN layers.

Mathematically, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, where the set of nodes and the set of edges are represented by $\mathcal{V}(|\mathcal{V}| = n)$ and $\mathcal{E}$ respectively. Each node in the graph consists of a self-loop (*i.e.*, $(v, v) \in \mathcal{E}, \forall v \in \mathcal{V}$). We introduce the matrix $X \in \mathbb{R}^{n \times k}$, which defines features associated with each node of the graph. Each row ($\in \mathbb{R}^k$) corresponds to a node and specifies its k-dimensional feature vector. We define the feature matrix X based on label encoding for the possible labels in the ontology. However, it can define any category of features.

The new d-dimensional node representation matrix $Z^{(1)} \in \mathbb{R}^{n \times d}$ for the graph in a single layer GCN is computed as follows:

$$Z^{(1)} = \varphi(\tilde{A}XW_0)$$

Here $W_0 \in \mathbb{R}^{k \times d}$ is the weight matrix for the first layer. φ represents some non-linear activation function, such as $ReLU, \varphi(z) = max(0, z)$ and d is a hyper-parameter. Let A and D represent the adjacency matrix and the diagonal matrix of the graph. Then, the normalized symmetry adjacency matrix of the graph is represented by $\tilde{A}$ and defined as $\tilde{A} = D^{-1/2}AD^{1/2}$. In general, node representation matrix for the $(i+1)^{th}$ layer is computed using the formula:

$$Z^{(i+1)} = \varphi(\tilde{A}Z^{(i)}W_l)$$

$Z^{(0)}$ is the initial feature matrix of the node, *i.e,* X.

In order to accommodate GCN for the user interest tracing task, we exploit the direct and all the transitive concepts used in a sub-user representation graph (cf. ⑤ in Fig. 1). We utilize the concept label encoding strategy in order to define the initial feature vector of each concept c as discussed in the following subsection. Further, we design a two-layer GCN architecture which is followed by a linear and a sigmoid layer. The architecture is nourished with the sub-user representation graph along with the concept label encoding associated with the concept present in the graph. This network shares the information among the nodes which are maximum two hops far from each other. We utilize the same aggregate function as advised in [18]. Multiple kinds of readout operations have been discussed in [43] to get the graph level representation. We utilize the arithmetic mean of all the nodes as the readout operation to get the sub-user representation. Our observations based on the initial experiments on the validation set show that a two-layer GCN accomplishes better results than a single-layer GCN. Further, incorporating more layers did not help in improving the prediction performance.

Concept Label Encoding

The user interest tracing model based on GCN receives the sub-user representation graph as input in order to identify the Doppelgänger. This graph is a DAG where each node represents one of the concepts. In order to get a better understanding of the similarity among the sub-users, providing a GCN only with the graph structure is not sufficient. Thus, we define a label encoding for each distinguished node, *i.e.*, concept. To this end, we utilize a one-hot encoding scheme. For example, while conducting experiments for the English language, we create a list of all the concepts present in the training set. Let $|c|$ be the total number of different concepts identified in the previous step. Then, we define a vector of dimension $|c|$ for each of the concepts where each position of the vector corresponds to one of the concepts. The entries of the concept vector are initialized with value "0" except only one position set to "1" (one-hot vector).

We provide the sub-user representation graph and the concept vector for each node of the graph as an input. This matrix acts as the initial feature matrix for the basic GCN model. As it can be observed, the basic GCN model provides equal weight to each concept published/edited by a sub-user. However, it may happen that a sub-user has published/edited some concepts more than once, which is lost in the current configuration.

5.3 Semantic User Interest Tracing (SUIT)

The σ_{SUIT} model attempts to address the shortcomings of the above-mentioned GCN model (σ_{GCN}) by incorporating the frequency information. As mentioned, σ_{GCN} does not grant any kind of weight to the concept. We derive the weight of a concept through the appearance frequency of the respective concept in the sub-user representation. We incorporate the frequency information in the one-hot concept vector by integrating an additional dimension (cf. ⑤ in Fig. 1). The last column of the feature matrix now represents the frequency of the respective concepts. Thus, σ_{SUIT} provides the label as well as the weight information for each of the concepts within the sub-user representation graph.

6 EXPERIMENTAL EVALUATION

We now explain the experimental settings. To this end, we introduce the experimental setup before presenting the experimental data set and results. We also present a sensitivity study and our findings.

6.1 Experimental Setup

The task of user interest tracing aims at identifying the same users based on their semantic representation of publishing/editing behavior towards the Web documents. In particular, we focus on exploiting the concepts of the Web documents to derive the Doppelgänger users. We develop various models in English, German, and French. For conducting the experiments, we need two pieces of information: a set of users along with their published/edited Web documents. Due to the availability of an ample amount of concepts associated with every document, we settle for a realistic setting to make it more impactful since the other category structure like the Wikipedia Category System is noisy and not handled systematically. Therefore, we exploit the WordNet concepts for the documents as mentioned in YAGO [25], totaling 68, 423.

Data Set Extraction

In order to perform the experiments, we aim at inspecting the set of users along with their published/edited Web documents. One, if not the most paramount source for this sort of information is Wikipedia. For our experiments, we utilize a subset of the Wikipedia encyclopedia and its associated user community. More precisely, we extracted all the Wikipedians (users) from the European Union[3] using the available category members identifier Wikipedia API[4]. Then, we extracted each user's contributions in English Wikipedia by exploiting the revision history of the users. To this end, we utilize the user contribution Wikipedia API[5] for the retrieval of the user revision history. It is worth mentioning that we focus on the main Wikipedia articles edited by a user for the experiments. For the same users, we extracted their revision history for French and German versions of Wikipedia, as well. In the current experiments, we utilize revision history as of March 23, 2021, for English and June 1, 2021, for German and French versions of Wikipedia. Not surprisingly, it is observable that the European users interested in the English version of Wikipedia aren't necessarily interested in other versions of Wikipedia. It is because the English version is

[3] https://en.wikipedia.org/wiki/Category:Wikipedians_in_the_European_Union
[4] https://www.mediawiki.org/wiki/API:Categorymembers
[5] https://www.mediawiki.org/wiki/API:Usercontribs

Language	#Users	#Average Articles Edited	#Median Articles Edited
English	5400	307.35	18
German	2097	253.59	5
French	1125	242.37	4

Table 1: Statistics of Edited Articles

more globalized and contains a massive amount of documents compared to the other languages. Table 1 represents the statistics about different users along with their contributions in the Wikipedia articles, which validates our observation.

Evaluation Dataset

Since there is no proper annotated dataset available for this task, we follow the approach mentioned in [5, 15] and adapt it. For that purpose, we split a user into J sub-users and randomly assign an equal number of distinguished documents to each sub-user (cf. Sec. 3). For example, for $|u|$ number of users, we have $(J * |u|)$ sub-users in the dataset. The value of J is set to 100 in the current scenario. Each sub-user, along with its direct, transitive, and/or inherent semantic relationships among the concepts, are given as input to the respective models. We randomly split each user in $(80 : 20)$ out of the 100 assigned sub-users. We utilize the 20% dataset for testing purposes. Further, we split the 80% dataset and utilize 90% of that as training and 10% for validation purposes. Here we should pay attention that for each set of J sub-users for a user u_i, we may have a very high number of documents associated with a sub-user, and only those J sub-users have an equal number of distinguished documents. It means that the total number of documents in the training set will not be exactly in the same ratio $(80 : 20)$ as for the test set. We maintain the ratio of $(80 : 20)$ for the total number of sub-users derived from all the users u. We repeat the same steps for all the languages. In order to predict user publishing/editing patterns, we performed extensive experiments with different thresholds of documents and settled ourselves to the users who have published text in at least 100 different documents, since less number of edited documents are not enough to generate the patterns. We report the detailed results for document thresholds of 500 and 1000 in Tables 3 and 4 (cf. Sec. 6.3) and left the other thresholds for the sensitivity study (cf. Sec. 6.4). The statistics for different document thresholds, along with the number of users and documents associated with the train and the test set, are shown in Table 2. For the sake of reproducibility, the dataset has been made publicly accessible at the project page of SUIT[6].

6.2 Model Configurations

For the random forest based models, we utilize the Scikit-learn library[7]. The bootstrap sample aggregation of the training data and the gini impurity measure to quantify the quality of a split have been exploited for the training of random forest based models. The number of decision trees in the forest is 100. We implement the GCN based models by exploiting PyTorch[8] and DGL[9] libraries. The pre-sigmoid logits are attained by operating the two hidden layers of convolution followed by a linear layer. A geometric pyramid rule [26] assigns the number of neurons in the respective convolutional layers. Both the GCN based models are trained by employing Adam optimization technique. The learning rate and the number of epochs are 0.001 and 100, respectively. After performing several experiments on the validation set with different settings, we identified the aforementioned configurations as best performing due to their better generalization.

6.3 Experimental Results

We conduct a wide range of experiments based on the previously discussed experimental settings. We develop several models for all three languages based on approaches described in Section 5. In order to show the difficulty of the task, we develop a "naive" baseline and call it random (σ_{Rand}). This method randomly selects a user from a set of candidate users and assigns it to a document. In all the models, we focus on the derived concepts associated with a document via a knowledge graph in order to create user publishing/editing patterns. Due to this, all the proposed models are semantic and language-agnostic in nature. Additionally, we also develop a baseline model (σ_{Base}). More specifically, (σ_{Base}) is based on "supervised authorship attribution" problem as in [2]. It solves the problem by employing stylometric features and a Support Vector Machine as a classifier. We include all the features except the "leetspeak" since our experiments are based on Wikipedia articles, and leetspeak (or Internet slang) is very uncommon in Wikipedia.

We summarize the macro-averaged and the micro-averaged scores for document thresholds of 500 and 1000 in Tables 3 and 4, respectively. One can observe that both the GCN based approaches dominate the random forest as well as the baseline models with a larger margin. In particular, σ_{SUIT} outperforms the competitor models by a margin of around 7% to 16% in all three languages for macro-averaged F-measure score. σ_{SUIT} beats σ_{Base} with a margin of at least 36% in macro-averaged F-score across all the languages. It can also be observed that baseline approaches have a very high precision value compared to the recall value. This gap between measures is because these models are able to capture the patterns for the highly active users. On the contrary, they fail to do for the less active ones. This difference can be observable both in macro and micro average scores. The excelling performance of the GCN based models is accredited to their capability of better encoding of the structured inherent semantic among the concepts. In addition, σ_{SUIT} performs significantly high, because of its conceptual adaptation of giving more weight to the more significant concept. Following the same line of observation, the σ_{SUIT} model

[6] https://spaniol.users.greyc.fr/research/SUIT/

[7] https://scikit-learn.org/

[8] https://pytorch.org/

[9] https://www.dgl.ai/

Language	Statistics	Document Threshold									
		100	200	300	400	500	600	700	800	900	1000
English	#Users	1,345	901	695	568	467	412	351	322	297	269
	Train Set	805,919	791,179	778,299	766,606	753,451	744,367	730,506	723,419	716,804	707555
	Test Set	267,431	259,720	253,259	247,621	241,518	237,377	231,577	228,457	225,449	221,609
German	#Users	281	215	170	147	126	121	111	102	95	83
	Train Set	309,203	305,687	301,253	297,664	293,341	292,163	289,493	286,454	283,720	278,841
	Test Set	93,583	92,181	90,389	89,062	87,513	87,097	86,068	84,954	83,961	82,083
French	#Users	133	101	80	71	65	60	56	51	50	48
	Train Set	171,773	169,475	166,647	165,127	163,721	162,581	161,433	159,568	159,041	158,120
	Test Set	49,461	48,714	47,778	47,262	46,796	46,362	45,910	45,266	45,110	44,784

Table 2: Statistics of Users, Train and Test Set with Different Document Threshold (Gold Standard)

also outperforms the other competitor models in micro-averaged F-measure score with a margin of 3% to 7% across all three languages. Here, the baseline model σ_{Base} is beaten by at least 25% in F-measure score. The micro-averaged score is relatively high compared to the macro-averaged one since some instances (more active users) are performing better, and the micro-averaged score is influenced by the documents associated with those active users. In contrast, the macro-averaged score treats each instance equally.

6.4 Sensitivity Analysis

In addition to the previously reported results, we also present a sensitivity study based on the different document thresholds. In particular, we analyze the performance of different approaches by altering the document thresholds value from 100 to 1000. A document threshold of p means that we develop models only for those users who have published texts in at least p different documents. The statistics of users with different thresholds are reported in Table 2. We illustrate the macro and micro averaged F-measure score with different document threshold across all the three languages in Fig. 2. As we can observe, the performance of all the models is generally increasing with the increment of the threshold value. GCN based approaches (green and blue dotted lines) dominate the other approaches across all languages. This supports our hypothesis that GCN is capable of representing the better encoding of the inherent semantic among the concepts. The green dotted line at the top of each graph claims the superiority of σ_{SUIT} among all the models and supports the hypothesis that significant concepts deserve more priority. The same pattern can also be observed for the document thresholds of 500 and 1000, which are reported in Tables 3 and 4 (cf. Sec. 6.3). The increasing performance of the models with the increment in the threshold value highlights that the patterns for the most active users (*i.e.*, users who are publishing texts frequently) are relatively easy to predict in comparison with the less active ones. This also enlightens that more coverage of concepts provides a better representation of the sub-user, and its versatility is more capable of identifying the patterns among the sub-users.

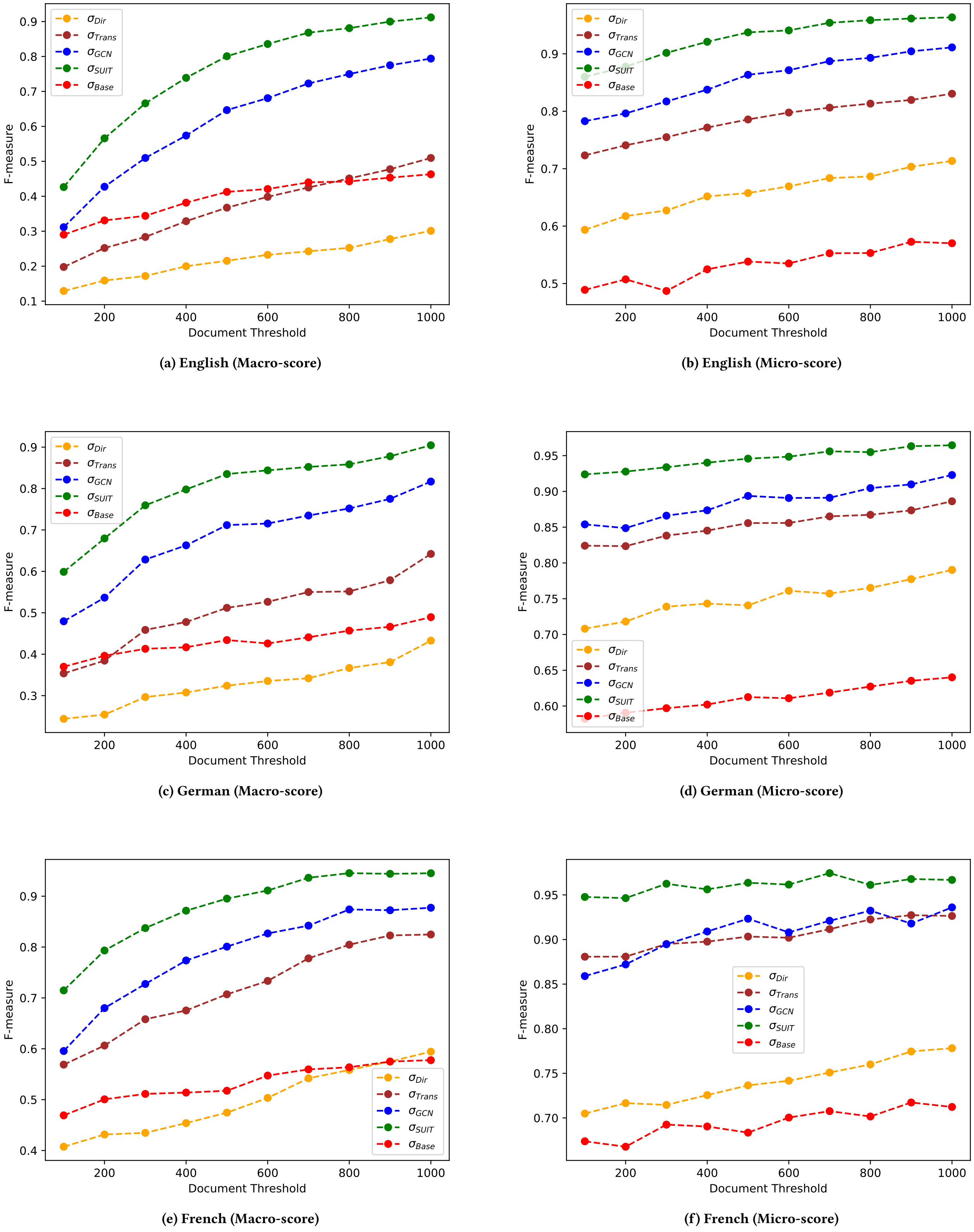
6.5 Findings

Our hypothesis that concepts are an excellent indicator to predict a user publishing/editing pattern was confirmed through extensive experiments in several languages. As discussed in Sections 3 - 5, the entire methodology is entirely based on concepts derived from an associated Web document. As our approach is purely semantic and - thus - language-agnostic in nature, it does not require specific linguistic features. This entails that our methodology is conceptually adaptable to any language. Moreover, we also noticed that incorporating the transitive concept information in prediction leads to a further improvement of the prediction model, as reported in Tables 3 and 4. The reason for this behavior can be attributed to the fact that the direct concepts only provide very focused information, whereas integration of transitive concepts allows the model to learn more facets as well as is able to generalize concept dependencies.

7 ETHICAL CONSIDERATIONS

Given the comprehensive experiments in the previous section (cf. Sec. 6.3 for details), we can conclude that it becomes effectively and efficiently possible to trace a user solely based on the semantic concepts he/she is interested in. As such, our study shows that employing abstraction of user interests by means of conceptualization supports a very fine-grained user modeling/tracing. In particular, by raising the pattern analysis to the entity-level, a very concise user (interest) tracing becomes possible. While existing user tracing approaches are mostly mono-lingual and/or mono-community, the here presented approach highlights a more sophisticated method by raising analytics and prediction to the entity-level. As a consequence, semantic user interest tracing (SUIT) provides a very powerful mean of user identification.

While our experiments have shown that due to our language-agnostic approach users become traceable across languages, there



(a) English (Macro-score)

(b) English (Micro-score)

(c) German (Macro-score)

(d) German (Micro-score)

(e) French (Macro-score)

(f) French (Micro-score)

Figure 2: Illustration of F-measure Scores for Different Document Thresholds

Language	Metrics	Macro-average						Micro-average					
		σ_{Rand}	σ_{Base}	σ_{Dir}	σ_{Trans}	σ_{GCN}	σ_{SUIT}	σ_{Rand}	σ_{Base}	σ_{Dir}	σ_{Trans}	σ_{GCN}	σ_{SUIT}
English	Precision	0.0027	0.5427	0.3317	0.5223	0.6762	**0.8119**	0.0028	0.5387	0.9986	**0.9993**	0.8756	0.9407
	Recall	0.0032	0.3328	0.1595	0.2834	0.6195	**0.7898**	0.0028	0.538	0.4902	0.6473	0.8517	**0.9342**
	F-measure	0.0029	0.4126	0.2154	0.3674	0.6466	**0.8007**	0.0028	0.5383	0.6576	0.7857	0.8635	**0.9374**
German	Precision	0.0061	0.525	0.4916	0.68	0.7672	**0.8552**	0.0055	0.6138	**0.9998**	0.9997	0.9225	0.9522
	Recall	0.0086	0.3702	0.2416	0.4105	0.6637	**0.8158**	0.0055	0.6113	0.5882	0.7479	0.8668	**0.9395**
	F-measure	0.0071	0.4342	0.324	0.512	0.7117	**0.835**	0.0055	0.6125	0.7407	0.8557	0.8937	**0.9458**
French	Precision	0.0208	0.6071	0.7067	**0.9538**	0.8462	0.9118	0.0192	0.6842	0.999	**1**	0.9483	0.9682
	Recall	0.0322	0.4509	0.3571	0.5617	0.7601	**0.8792**	0.0192	0.6828	0.5832	0.8236	0.8996	**0.9591**
	F-measure	0.0253	0.5175	0.4745	0.7071	0.8008	**0.8952**	0.0192	0.6835	0.7365	0.9033	0.9233	**0.9636**

Table 3: Macro- and Micro-average Results for Document Threshold of 500

Language	Metrics	Macro-average						Micro-average					
		σ_{Rand}	σ_{Base}	σ_{Dir}	σ_{Trans}	σ_{GCN}	σ_{SUIT}	σ_{Rand}	σ_{Base}	σ_{Dir}	σ_{Trans}	σ_{GCN}	σ_{SUIT}
English	Precision	0.0021	0.5846	0.451	0.689	0.8082	**0.919**	0.002	0.5718	0.9989	**0.9995**	0.9161	0.9653
	Recall	0.0033	0.3832	0.226	0.4044	0.7802	**0.9049**	0.002	0.5686	0.555	0.7104	0.9064	**0.9617**
	F-measure	0.0025	0.463	0.3011	0.5096	0.794	**0.9119**	0.002	0.5702	0.7135	0.8305	0.9112	**0.9635**
German	Precision	0.0068	0.5535	0.6145	0.8313	0.8502	**0.9184**	0.0083	0.6413	**1**	**1**	0.9415	0.9684
	Recall	0.0116	0.4387	0.3341	0.5232	0.7859	**0.8911**	0.0083	0.6389	0.6533	0.7956	0.9049	**0.9606**
	F-measure	0.0085	0.4895	0.4328	0.6422	0.8168	**0.9045**	0.0083	0.6401	0.7903	0.8862	0.9229	**0.9645**
French	Precision	0.0179	0.6496	0.8112	**0.9792**	0.9182	0.9529	0.0171	0.7128	0.999	**1**	0.9626	0.9703
	Recall	0.025	0.52	0.4689	0.7124	0.84	**0.9374**	0.0171	0.7117	0.6373	0.8629	0.9109	**0.9632**
	F-measure	0.0209	0.5776	0.5943	0.8247	0.8773	**0.9451**	0.0171	0.7123	0.7782	0.9264	0.936	**0.9668**

Table 4: Macro- and Micro-average Results for Document Threshold of 1000

is also a huge potential of exploiting semantic user interest tracing across communities. However, at the same time, our observations highlight a potential privacy issue and raise the question: "To what extend Web users should be profiled in order to preserve a balance between personalization and surveillance?". In times of increasing repression of Web users and limitation of free speech in many countries across the globe, we - therefore - argue that further and more sophisticated studies are needed in order to analyze the impact of (entity-level) personalization in cross-lingual as well as in cross-community settings. In particular, we claim that there is a need in raising awareness of the inherent surveillance risks and protecting the average Web user's privacy. For that purpose, automatic assessment methods should be investigated that might intentionally inject "arbitrary" concepts into a Web user's profile.

8 CONCLUSIONS & FUTURE WORK

In this paper, we presented a novel method of semantic user interest tracing called SUIT. To the best of our knowledge, our approach is unique by entirely building on (semantic) concepts in order to trace user interests and analyze/predict their publishing/editing behavior. To this end, we adapted a GCN by defining concept label encodings and node weight information within the concept relationship graph of a sub-user. In our comprehensive studies, we performed experiments in multiple languages. Here, we have highlighted that SUIT outperforms state-of-the-art approaches, including a baseline GCN implementation. In particular, the enhanced performance of the SUIT model for German and French demonstrate the viability of our method also for languages with less ample resources. As a result, user interest traces can be effectively and efficiently revealed. However, this raises at the same time serious risks of a potential privacy intrusion, particularly, because of the language-agnostic nature of our approach.

In future work, we intend to pursue at least two more studies. First, we aim at looking into privacy aspects of user interest tracing beyond communities and social networks. To this end, we will study semantic user interest tracing across social media, such as Twitter or

Facebook and news article comment sections. Since our presented approach is language agnostic, we also aim at covering contents in multiple languages. Second, we plan to extend our study in order to target the user credibility, too. For that purpose, we will examine the applicability of SUIT in the context of fake news detection, in particular, the identification of "semantically suspicious" user publication patterns.

REFERENCES

[1] Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. *ACM Trans. Inf. Syst.* 26, 2, Article 7 (April 2008), 29 pages.

[2] Sadia Afroz, A. Caliskan-Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger Finder: Taking Stylometry to the Underground. *Proceedings - IEEE Symposium on Security and Privacy* (11 2014), 212–226. https://doi.org/10.1109/SP.2014.21

[3] Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb, and Abdelmajid Ben Hamadou. 2016. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. *Applied Intelligence* 45 (09 2016). https://doi.org/10.1007/s10489-015-0755-x

[4] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32.

[5] Despoina Chatzakou, Juan Soler-Company, Theodora Tsikrika, Leo Wanner, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2020. User Identity Linkage in Social Media Using Linguistic and Social Interaction Features. In *12th ACM Conference on Web Science* (Southampton, United Kingdom) *(WebSci '20)*. Association for Computing Machinery, New York, NY, USA, 295–304.

[6] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. 2010. *Short and Tweet: Experiments on Recommending Content from Information Streams*. Association for Computing Machinery, New York, NY, USA, 1185–1194.

[7] Marco Cristani, Giorgio Roffo, Cristina Segalin, Loris Bazzani, Alessandro Vinciarelli, and Vittorio Murino. 2012. Conversationally-Inspired Stylometric Features for Authorship Attribution in Instant Messaging. In *Proceedings of the 20th ACM International Conference on Multimedia* (Nara, Japan) *(MM '12)*. Association for Computing Machinery, New York, NY, USA, 1121–1124.

[8] Alia El Bolock, Cornelia Herbert, and Slim Abdennadher. 2020. CCOnto: Towards an Ontology-Based Model for Character Computing. In *Research Challenges in Information Science*. Springer International Publishing, Cham, 529–535.

[9] Zakaria Elberrichi, Abdellatif Rahmoun, and Mohamed Bentaallah. 2008. Using WordNet for Text Categorization. *Int. Arab J. Inf. Technol.* 5 (01 2008), 16–24.

[10] Daniela Godoy and Analía Amandi. 2006. Modeling User Interests by Conceptual Clustering. *Inf. Syst.* 31, 4–5 (June 2006), 247–265.

[11] Zhiwei Guo and Heng Wang. 2021. A Deep Graph Neural Network-Based Mechanism for Social Recommendations. *IEEE Transactions on Industrial Informatics* 17, 4 (2021), 2776–2783. https://doi.org/10.1109/TII.2020.2986316

[12] Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto. 2017. Knowledge Transfer for Out-of-Knowledge-Base Entities : A Graph Neural Network Approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 1802–1808.

[13] Jonghyun Han and Hyunju Lee. 2014. Characterizing User Interest Using Heterogeneous Media. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14 Companion)*. Association for Computing Machinery, New York, NY, USA, 289–290.

[14] Jonghyun Han and Hyunju Lee. 2016. Characterizing the Interests of Social Media Users. *Inf. Sci.* 358, C (Sept. 2016), 112–128.

[15] F. Johansson, Lisa Kaati, and Amendra Shrestha. 2015. Timeprints for identifying social media users with multiple aliases. *Security Informatics* 4 (2015), 1–11.

[16] Jaeyong Kang, Hongseok Choi, and Hyunju Lee. 2018. Deep recurrent convolutional networks for inferring user interests from social media. *Journal of Intelligent Information Systems* 52 (2018), 191–209.

[17] Pavan Kapanipathi, Prateek Jain, Chitra Venkatramani, and A. Sheth. 2014. User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In *11th Extended Semantic Web Conference (ESWC)*. 99–113.

[18] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR 2017, April 24-26, Toulon, France.*

[19] Niklas Kolbe, Pierre-Yves Vandenbussche, Sylvain Kubler, and Yves Le Traon. 2020. LOVBench: Ontology Ranking Benchmark. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. ACM, New York, NY, USA, 1750–1760.

[20] Amit Kumar, Govind, Céline Alec, and Marc Spaniol. 2020. Blogger or President? Exploitation of Patterns in Entity Type Graphs for Representative Entity Type Classification. In *Proc. of the 12th Intl. ACM Web Science Conference (WebSci '20)*. 59–68.

[21] Amit Kumar, Govind, and Marc Spaniol. 2021. Semantic Search via Entity-Types: The SEMANNOREX Framework. In *Companion Proceedings of the Web Conference 2021 (WWW2021)*. ACM, New York, NY, USA, 690–694.

[22] Yongjun Li, Zhaoting Su, Jiaqi Yang, and Congjie Gao. 2020. Exploiting similarities of user friendship networks across social networks for user identification. *Information Sciences* 506 (2020), 78–98.

[23] Zhiwei Liu, Yingtong Dou, Philip S. Yu, Yutong Deng, and Hao Peng. 2020. Alleviating the Inconsistency Problem of Applying Graph Neural Network to Fraud Detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1569–1572.

[24] Vasyl Lytvyn, Victoria Vysotska, Oleh Veres, Ihor Rishnyak, and Halya Rishnyak. 2017. *Classification Methods of Text Documents Using Ontology Based Approach*. Springer International Publishing, Cham, 229–240.

[25] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *7th Biennial Conference on Innovative Data Systems Research (CIDR 2015), January 4-7, 2015, California, USA.*

[26] T. Masters. 1993. *Practical neural network recipes in C++*. Morgan Kaufmann.

[27] Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015. Exposing Paid Opinion Manipulation Trolls. In *Proceedings of the International Conference Recent Advances in NLP*. Hissar, Bulgaria, 443–450.

[28] R. Ottoni, D.L. Casas, J. P. Pesce, Wagner Meira Jr, C. Wilson, A. Mislove, and Virgilio Almeida. 2014. Of pins and tweets: Investigating how users behave across image-and text-based social networks. *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014)* (2014), 386–395.

[29] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. 2020. Doctor XAI: An Ontology-Based Approach to Black-Box Sequential Data Classification Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. ACM, New York, NY, USA, 629–639.

[30] Krishnan Ramanathan and Komal Kapoor. 2009. Creating User Profiles Using Wikipedia. In *Conceptual Modeling - ER 2009*. Springer Berlin Heidelberg, Berlin, Heidelberg, 415–427.

[31] Andi Rexha, S. Klampfl, Mark Kröll, and Roman Kern. 2015. Towards Authorship Attribution for Bibliometrics using Stylometric Features. In *CLBib@ISSI*.

[32] Andi Rexha, S. Klampfl, Mark Kröll, and Roman Kern. 2016. Towards a More Fine Grained Analysis of Scientific Authorship: Predicting the Number of Authors Using Stylometric Features. In *BIR@ECIR*.

[33] Andi Rexha, Mark Kröll, Hermann Ziak, and Roman Kern. 2018. Authorship identification of documents with high content similarity. *Scientometrics* 115 (2018), 223 – 237.

[34] Weijing Shi and R. Rajkumar. 2020. Point-GNN: Graph Neural Network for 3D Object Detection in a Point Cloud. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 1708–1716.

[35] Rui Sousa Silva, Gustavo Laboreiro, Luís Sarmento, Tim Grant, Eugénio Oliveira, and Belinda Maia. 2011. 'twazn me!!! ;(' Automatic Authorship Analysis of Micro-Blogging Messages. In *Natural Language Processing and Information Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 161–168.

[36] E. Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Assoc. Inf. Sci. Technol.* 60 (2009), 538–556.

[37] Alvin Toffler. 1980. *The third wave / by Alvin Toffler* (1st ed. ed.). Morrow New York. 544 p. ; pages.

[38] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJXMpikCZ

[39] Esther Villar-Rodriguez, Javier Del Ser, Miren Nekane Bilbao, and Sancho Salcedo-Sanz. 2016. A feature selection method for author identification in interactive communications based on supervised learning and language typicality. *Engineering Applications of Artificial Intelligence* 56 (2016), 175–184. https://doi.org/10.1016/j.engappai.2016.09.004

[40] Xiaoyang Wang, Yao Ma, Yiqi Wang, Wei Jin, Xin Wang, Jiliang Tang, Caiyan Jia, and Jian Yu. 2020. Traffic Flow Prediction via Spatial Temporal Graph Neural Network. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1082–1092.

[41] Zhe Wang, Chun-Hua Wu, Qing-Biao Li, Bo Yan, and Kang-Feng Zheng. 2020. Encoding Text Information with Graph Convolutional Networks for Personality Recognition. *Applied Sciences* 10, 12 (2020). https://doi.org/10.3390/app10124081

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

[43] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, C. Zhang, and Philip S. Yu. 2021. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 32 (2021), 4–24.

[44] 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* 14, 5 (2020), 445–617. https://doi.org/10.1561/1500000078

[45] Chunxia Zhang, Xindong Wu, Zhendong Niu, and Wei Ding. 2014. Authorship identification from unstructured texts. *Knowledge-Based Systems* 66 (2014), 99–111. https://doi.org/10.1016/j.knosys.2014.04.025

[46] Shijie Zhang, Hongzhi Yin, Tong Chen, Quoc Viet Nguyen Hung, Zi Huang, and Lizhen Cui. 2020. GCN-Based User Representation Learning for Unifying Robust Recommendation and Fraudster Detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. ACM, New York, NY, USA, 689–698.

[47] Fan Zhou, Xiuxiu Qi, Xovee Xu, Jiahao Wang, Ting Zhong, and Goce Trajcevski. 2020. Meta-Learned User Preference for Topic Participation Prediction. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. 1–6.