On the Quality of Semantic Interest Profiles for Online Social Network Consumers

Christoph Besel
University of Passau
Innstraße 41
Passau, Germany
christoph.besel@googlemail.com

Jörg Schlötterer
University of Passau
Innstraße 41
Passau, Germany
joerg.schloetterer@uni-passau.de

Michael Granitzer
University of Passau
Innstraße 41
Passau, Germany
michael.granitzer@uni-passau.de

ABSTRACT

Social media based recommendation systems infer users' interests and preferences from their social network activity in order to provide personalised recommendations. Typically, the user profiles are generated by analysing the users' posts or tweets. However, there might be a significant difference between what a user *produces* and what she *consumes*. We propose an approach for inferring user interests from followees (the accounts the user follows) rather than tweets. This is done by extracting named entities from a user's followees using the English Wikipedia as knowledge base and regarding them as interests. Afterwards, a spreading activation algorithm is performed on a Wikipedia category taxonomy to aggregate the various interests to a more abstract and broader interest profile. We evaluate the coverage of followee lists in terms of named entities and show that they provide sufficient input to infer comprehensive semantic interest profiles. Further, we compare the profiles created with the followee-based approach against tweet-based profiles. With over 7 out of 10 items being relevant to the users in our evaluation, we show that the followee-based approach can compete with the state of the art and performs even better in predicting the users' interests than their human friends do.

CCS Concepts

•Information systems → Personalization; •Human-centered computing → *Social networks;*

Keywords

Personalization, Twitter User Profile

1. INTRODUCTION

We have seen a rapid increase in the amount of published information and data since the rise of the Internet. Obviously, it is not possible for humans to process all the information available, a problem known as "information overload" [3]. At the same time more and more people reveal their interests explicitly in and implicitly by using social networks. The goal of social media based recommendation systems is to infer users' interests and preferences from their social network activity and use the thereby generated interest profiles for making personalized content recommendations. Using social information for recommendation systems is also connected to the hope of solving the cold start problem which in particular correlation based approaches suffer from, especially for smaller web pages. The cold start problem concerns the issue that a system does not know anything about new users and needs an initial phase to gather information about them.

Most of the related work infers the interest profiles from a user's posts or tweets. However, there might be a significant difference between what a user *produces* and what she *consumes*. Moreover the passive use of social network sites is on the rise. Now four in ten users browse Facebook only passively, without posting anything [6]. For those users, profile construction based on a user's postings fails, since there is simply no input from which the profile could be created. We address this problem by inferring semantic interest profiles from the twitter followees (the accounts, the user follows) rather than her tweets. It is to note, that while we focus on Twitter and followees, the approach could be adapted to other online social networks as well, by accounting for the corresponding features, e.g. *likes* on Facebook.

The rationale for the followee-based approach is that many famous people maintain a Twitter account and a lot of Twitter users follow these accounts. For those accounts, the likelihood that a Wikipedia article about this person exists is very high. Moreover, Wikipedia articles are typically linked to higher level categories (e.g., the article about the football player "Thomas Müller" is linked to the category "German footballers"). Making use of those categories, following an account that can be linked to a Wikipedia article can be seen as implicit expression of interests (e.g., following the football player "Thomas Müller" reveals interest in "German footballers"). In addition, the assigned categories are organised in some kind of hierarchy in Wikipedia, thus they can be traversed in order to provide a more fine- or coarse-grained profile. This approach immediately raises the question of whether a sufficient number of followees can be linked to Wikipedia entities, which we address in the first part of the paper.

Specifically, the contribution of this paper is the following:

- We evaluate the coverage of followee lists in terms of named entities in the English Wikipedia and show that the followee lists provide enough input to infer comprehensive semantic interest profiles.
- We propose a followee-based approach to create user interest profiles, which can compete with state of the art tweet-based approaches.
- We compare the similarity of followee- and tweet-based profiles and show that they are more similar on very concrete and abstract levels than in between.

The remainder of the paper is organized as follows: In the next section, we present related work in the field of social media based recommendation systems. Then we provide an overview of the approach, followed by the evaluation of named entity coverage in followee lists and the evaluation of the overall quality of the approach by a user study. In the last part, we compare and analyze the similarity of profiles generated with the proposed approach against tweet-based profiles. Finally, we conclude the paper and provide an outlook on future work.

2. RELATED WORK

Research on user profiling and personalized content recommendation has been done for many years since the beginning of the web [10]. Early approaches focused on the web [10, 9] and search history [18] of the user. Recently, with the emergence of social networks like Twitter, research has shifted to analyze user activities on these platforms. For instance Siehndel and Kawase [17] introduced *TwikiMe*, a prototype for generating user profiles by extracting entities from the user's tweets and linking them to the 23 top-level categories of the English Wikipedia. This leads to abstract interest profiles with a fixed size represented as a 23-length vector.

Abel et. al. [1] in their work compared hastag-based, topic-based (bag-of-words) and entity-based user models generated from the user's tweets, for news recommendation. In this approach the scoring of the extracted concepts and interests is based on a simple term frequency technique. The results of their comparative evaluation showed that the simple bag-of-words and hashtag-based approaches, which did not consider the semantics of a tweet, were clearly outperformed by the (semantic) entity-based strategy (precision of 0.71 compared to 0.4 and 0.1). Based on these results Tao et. al. [19] presented *TUMS*, a Twitter-based User Modeling Service, that tries to infer semantic user profiles from the messages people post on Twitter. However, the focus of *TUMS* is to make use of semantic web technologies for providing a standardized representation of the interest profiles allowing an easy exchange between different web services. This is connected with the hope to solve the so-called ramp up or cold start problem, a downside of approaches like content based or collaborative filtering [19, 12], which usually depend on the build-up of a user history before making personalized content recommendations. In terms of the applied algorithm and the knowledge base, the approach introduced by Kapnipathi et. al [7] is the closest to our work. They used the English Wikipedia to spot entities in tweets and leveraged the hierarchical relationships by performing a spreading activation on the Wikipedia Category Graph to infer user interests. The result, a weighted hierarchical interest profile (expressed as a so-called *Hierarchical Interest Graph*), was evaluated by a user study which showed an average of approximately eight out of the ten interests in the graph being relevant to a user.

Even though Siehndel and Kawase [17] suggested investigating other types of inputs for inferring user interests, most of the related work only makes use of the content posted by a user (e.g. the tweets). Some approaches tried to consider the social graph of the user at least to some extent [14, 12] whereas Lim and Datta [11] presented a basic approach for interest profile creation based on celebrities, a user follows. These celebrities are classified as belonging to one or more of 15 predefined interest categories. The classification is based on the celebrity's *occupation* field on his or her Wikipedia page and a set of keywords associated with each interest category. While this approach is also based on followees, in contrast to our work, it ignores the category information provided by Wikipedia and provides only support for a fixed (and predefined) set of interests, similar to Siehndel and Kawase [17]

Most similar to our work, a recent approach by Faralli et al. [4] also utilizes followees and the Wikipedia Bitaxonomy. However, while there are similarities in the applied approach, Faralli et al. do not directly evaluate the semantic interest profiles. Instead, they use it to identify users as belonging to a target population or not. Further, they apply itemset mining and based on the itemsets and association rules, they provide recommendations, e.g. for topical friends or categories a user might also be interested in and evaluate those recommendations. We in contrast evaluate, whether the constructed profile really describes the user.

3. APPROACH OVERVIEW

The generation of interest profiles in this paper can be seen as a four-step process which is shown in fig. 1. In the following, each step is described in more detail and a fictional user called *@soccerfan* will be used as an illustrating example.

Fetch user's friends In the first step the accounts which are followed by the user (the followees) are crawled. This is done through Twitter's RESTful Web API[1]. As the API applies strict rate limits, extensive use of caching techniques is made to reduce the number of requests sent to Twitter.

The fictional user *@soccerfan* might, among others, follow the accounts *@Cristiano (Cristiano Ronaldo)*, *@BSchweinsteiger (Basti Schweinsteiger)*, *@neymarjr (Neymar Jr)*, *@esmuellert_ (Thomas Müller)* and *@FIFAcom (FIFA.com)*.

Link friends to entities The objective of this step is to link the user's followees to corresponding entities represented by Wikipedia articles. This entity linking includes handling coincidental homonymy and ambiguity (for instance there are several famous "Thomas Müllers" with their own Wikipedia page). For that

[1] `https://dev.twitter.com/rest/public`

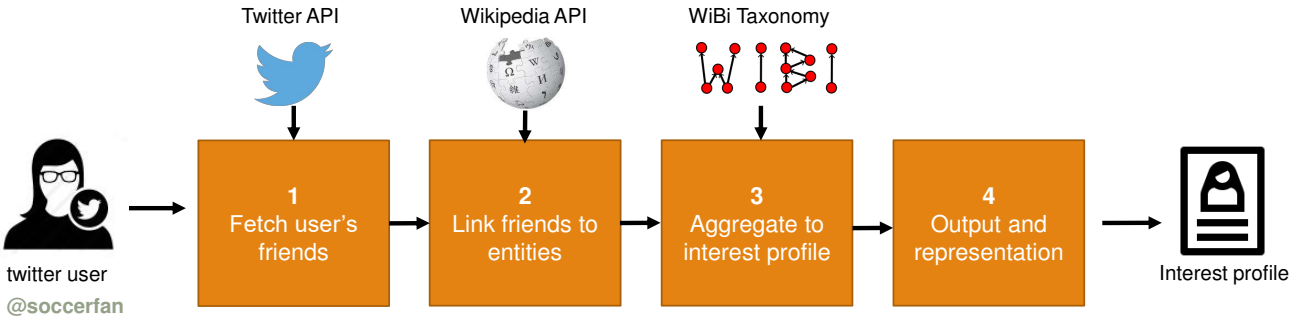

Figure 1: Overview approach

purpose the MediaWiki Web API[2] is used and several disambiguation heuristics are applied. They include syntactical measures (overlap coefficient of last 20 tweets and article summary) and probabilistic heuristics (Sense Prior and a reverse linking of Wikipedia articles to Twitter search results).

In our example the following entities might be extracted: *WikipediaPage:Christiano Ronaldo*, *WikipediaPage:Bastian Schweinsteiger* and *WikipediaPage:Thomas Mueller (footballer)*. As you can see, "Thomas Müller" was correctly linked to the famous football player.

Aggregate to interest profile The extracted Wikipedia article entities are assigned to Wikipedia categories. These categories are hierarchically structured (at least to some extent) and used to represent particular interests of the user. By performing a spreading activation algorithm on the Wikipedia Bitaxonomy (a taxonomy based on the Wikipedia page and category hierarchy [5]) the single interest entities are aggregated to a more abstract and broader interest profile.

The categories of the Wikipedia page entities extracted in the previous step represent the set of initially activated nodes. Their activation is spread during several iterations to neighboring nodes connected by outgoing edges. Formally the activation $a(v)$ of a node v can be written as:

$$a_t(j) \leftarrow a_{t-1}(j) + d \cdot a_{t-1}(i) \qquad (1)$$

where j is being activated by node i and $0 < d < 1$ represents the decay factor. If a node is activated by more than one node the activation is accumulated in this node. Apart from that, a normalization with the number of incoming edges and a so-called Intersection Boost (see [7] for more details), boosting nodes that are intersections of different paths are applied.

In our example the entities (pages) are assigned to categories such as *2014 FIFA World Cup players* or *German footballers*. Performing spreading activation identifies *sports* and *footballers* as two of the most suitable overall interest categories for the example user.

[2] `https://www.mediawiki.org/wiki/API`

Output and representation As the output of step three is a graph data structure with weighted nodes, the objective of this last step is to convert this representation to a common exchange format. Therefore, the top-k interests are extracted and can be represented in an arbitrary format. Typical representations include JSON or XML and semantic web vocabularies, such as the *FOAF*[3] (Friend of a Friend) or *Weighted Interests Vocabulary*[4] could be used. This also allows the provision of the interest profiles to other applications and web services through standardized interfaces.

4. ENTITY COVERAGE EVALUATION

The first question we need to address is whether the followee list of a Twitter user is sufficient input for inferring his or her interest profile. This mainly depends on the number of followees which could be linked to an entity and the quality of that entity linking. We evaluated both issues on a sample dataset.

4.1 Method and sample description

We conducted experimental research by crawling the profiles of 3000 twitter accounts (with over 350 000 followees in total) chosen randomly from an updated data set based on [2, 8]. Afterwards we analyzed the number of followees that could be linked to an entity and assessed the quality of that entity linking by applying the disambiguation heuristics mentioned in the second step of section 3.

A first analysis of the sample showed that over 72 % of the users in the sample are friends with more than 50 other accounts. More than half of the Twitter accounts examined had between 50 and 200 followees. The overwhelming majority (91 %) used the English language version of Twitter.

4.2 Quantitative results

For analyzing the number of followees that could be linked to a corresponding Wikipedia page entity we used the MediaWiki Web API[2]. As this API allows search on the English Wikipedia with an auto suggest feature enabled or disabled, we did the calculation for both. Table 1 and table 2 show

[3] `http://xmlns.com/foaf/spec/`

[4] `http://smiy.sourceforge.net/wi/versions/20100812/spec/weightedinterests.html`

the results for different selections on the sample. The numbers include the shares of followees which could be linked to an entity unambiguously, the followees that could be linked to more than one page (ambiguity) and the followees that could not be linked to any entity at all.

Table 1: Quantitative results (auto suggest enabled)

Selection	followees in % linked		
	unambiguously	ambiguously	not at all
None	69.89	7.14	22.72
Number of followees > 50	71,08	7.11	21.65
Number of followees < 50	66.77	7.23	25.51
English language version	71.24	7.24	21.27
Other language version	54.84	6.05	38.87
English language version, number of followees > 50	72.44	7.20	20.22

On average about 70 % of the total number of followees could be linked unambiguously to an entity by the MediaWiki API with the auto suggest feature enabled. In less than every tenth case (7.14 %) more than one disambiguation (articles of the same name) was possible. About a fifth of the followees could not be linked to any entity even with the auto suggest feature enabled. Considering only accounts using the English language version the share of followees linked unambiguously is significantly higher (71.24 %) than with other language versions (54.84 %). The same effect, even though to a lesser extent, can be seen when comparing accounts that have more and less than 50 followees. The best success rate (72.44 %) is achieved by a combined selection of accounts using the English language version of Twitter with more than 50 followees.

Table 2: Quantitative results (auto suggest disabled)

Selection	followees in % linked		
	unambiguously	ambiguously	not at all
None	41.23	5.73	52.93
Number of followees > 50	42.74	5.81	51.35
Number of followees < 50	37.24	5.54	57.08
English language version	42.61	5.88	51.39
Other language version	25.84	4.06	69.99
English language version, number of followees > 50	44.17	5.95	49.79

Table 3: Qualitative results (overlap coefficient)

	Entity Linking $n = 7500$		Baseline $n = 7500$	
	M	SD	M	SD
no normalization	0.2325	0.0910	0.2168	0.0871
normalization	0.0609	0.0643	0.0369	0.0365

M: mean, SD: standard deviation

With auto suggest feature disabled the share of followees that could be linked to an entity is, as one could expect, lower (41.23 % compared to 69.89 %). However the trends for the different sections are very similar. For accounts with more than 50 followees that use the English language version barely half could be linked to an entity (6 % of these ambiguously).

4.3 Qualitative results

The quantitative results may not necessarily imply that the quality of the entity linking is sufficient. This depends on whether the followee was linked with the semantically correct entity. For instance "common" people that share the name with a celebrity coincidentally might be linked to a Wikipedia page. To assess the quality of the entity linking we applied some of the disambiguation heuristics mentioned in section 3:

4.3.1 Overlap coefficient

Even though the applicability for tweets might be limited due to their short length and informal character we first calculated the overlap coefficient as a simple syntactic measure for assessing the link quality. This was done by collecting the last 20 tweets for 7500 randomly chosen Twitter users that could be linked to a Wikipedia page by the MediaWiki Web API[2] (auto suggest enabled) and the summary of the linked page (usually the very first section). Afterwards we tokenized the crawled input and converted it into a set of words, which also removed duplicates. On that basis we calculated the overlap coefficient as shown in eq. (2) (where X and Y are the two word token sets compared).

$$overlap(X, Y) = \frac{|X \cap Y|}{min(|X|, |Y|)} \tag{2}$$

The overlap coefficient was calculated for both, a random mapping of tweets and page summaries (the baseline) and for the linking suggested by the MediaWiki Web API. This was done before a text normalization (stop word removal and stemming) was applied as wells as afterwards. With text normalization the results show (see table 3) that the mean overlap coefficient for the entity linking is twice as high as for the random baseline mapping (Cohen's $d = 0.47$). Without text normalization the effect (Cohen's $d = 0.18$) is clearly smaller. All value differences were highly statistically significant ($p < 0.001$).

4.3.2 Reverse Linking

A very easy way to get a quick estimation of the entity linking quality is to search on Twitter for accounts with the name of the Wikipedia page title (entity). By doing this reverse linking we ended up in about 80 % of the cases with the account we started the entity linking from. Both Twitter and Wikipedia have optimized search indices and the article title sometimes contains additional disambiguation information (e.g. "footballer" for "Thomas Müller"). A high success rate could be seen as indicator of a good entity linking, but as the search algorithms and indices of Twitter and Wikipedia are black boxes for us this could only be a first clue.

4.3.3 Sense Prior

Sense Prior is a probabilistic approach which assumes that the most frequent word meaning dominates the others [16]. For that purpose the relative frequencies of so-called surface forms linking to an entity are calculated and the most frequent one is assumed to be the correct disambiguation. In this evaluation we used a dataset based on the internal link structure of the English Wikipedia to calculate the frequencies. Let $l = (s, a)$ be an internal link which points to article a with the link text (surface form) s and let $n(s)$ describe the number of occurrences of that surface form in all articles then

$$P(a|s) = \frac{l(s,a)}{n(s)} \tag{3}$$

is the probability that entity a is the correct disambiguation for surface form s.

We calculated that probability for 10 000 randomly chosen followee names (the surface form in this case) of our sample (no auto suggest, English language version and more than 50 followees) and compared the entity with the highest probability to the linked entity. If the Sense Prior dataset could provide a disambiguation (the case in 78 %) it corresponded with a probability of over 90 % with the linked entity.

4.4 Analysis and Discussion

The results of our empirical research show that without auto suggest almost half of the followees and with auto suggest over two thirds of the accounts a user is following could be linked to Wikipedia page entities successfully. This implies that the Twitter followees of a user actually could be a sufficient and broad basis for inferring interest profiles. As ambiguity does occur only in about one out of ten cases it should have little effect.
The qualitative evaluation points towards the same direction: With an overlap coefficient twice as high as for a random baseline mapping and success rates of about 80% for the reverse linking and around 80 % and 90 % for the probabilistic disambiguation heuristics the entity linking quality could be seen as sufficient as well.

We conclude, that the Twitter followees of a user provide already a sufficient input, both quantitatively and qualitatively, for inferring meaningful interest profiles.

5. USER STUDY

Even though the groundwork in the last section showed that the Twitter followees are a sufficient base for inferring interest profiles, the evaluation of personalization and/or recommendation systems typically involves a user study [7].
For that purpose we implemented the approach presented in section 3 in Python and evaluated it with real users. The modular application made use of several external modules such as *tweepy*[5] and *wikipedia*[6] for accessing the Web APIs and *networkx*[7] for building the taxonomy graph and performing spreading activation on it. The source code of the application can be found on the project repository[8].

5.1 Experimental Setup

For our evaluation we generated four different profile types defined by the number of iterations, the decay factor and the application of disambiguation heuristics (see table 4).

Table 4: Evaluated profile types

	Iterations	Decay	Disambiguation
Profile type 1	5	0.2	No
Profile type 2	5	0.2	Yes
Recommendations	3	0.2	Yes
Comparative Eval.	5	0.2	Yes

After the users had registered by providing their Twitter screenname and e-mail, they were notified by a mail providing a link to their personalized questionnaire. This questionnaire had four pages that corresponded with the four different interest profile types shown in table 4. For screenshots of the registration form and questionnaire pages please see the project repository[8].

On the first page the user was presented the top 20 interest categories (most weighted nodes) of the first profile type. The participants were asked to indicate their strength of interest for each category on a four-point Likert scale ranging from "very interesting" to "not interesting at all".
The second page was pretty much the same presenting the top 20 interests of profile type 2 that mainly differed in whether disambiguation heuristics were applied or not.
On the third page five Wikipedia articles that were assigned to the interest categories of the third interest profile (a smaller number of iterations was used to get more specific results) were shown. Again the participants were asked to indicate their strength of interest in the topics covered by these articles.
The last page showed the users ten interest categories randomly picked from the profiles of other users. As the categories did not appear in their interest profiles, no interest

[5] `http://www.tweepy.org/`
[6] `https://pypi.python.org/pypi/wikipedia/`
[7] `https://networkx.github.io/`
[8] The project repository includes application source code, evaluation scripts and screenshots `https://bitbucket.org/beselch/interest_twitter_acmsac16`

of the users in these categories was assumed and they were asked to evaluate whether this was correct or not.

Afterwards the participants were provided with a link they were asked to send a friend of theirs. This link lead to a one-paged survey that presented the user's friend with 20 interest categories. One half consisted of the top 10 interests of profile type 2 and the other half were the randomly picked interest categories that our approach assumed to be not interesting. Now the user's friend was asked to evaluate the interest of his or her friend in these interest categories. Following [20] these answers were used to compare the performance of the friend and our algorithm in predicting the user's interests. Whereas the pages one and two were obligatory the last two steps could be skipped by the participants.

5.2 Sample description

During the evaluation period from 30 June to 10 July 2015 64 Twitter users registered for the user study and 52 of them completed the survey (response rate of 81.25 %). A participant had 205 followees on average, while the median (114) was considerably lower. The used Twitter language versions were half German and half English. Barely half of the users posted fewer than 100 tweets (over 15 % nothing), which means that approaches based on the tweets would fail to generate interest profiles for that users. 46 participants submitted the optional third page and 17 people took part in the fourth step (comparative evaluation).

5.3 Results

5.3.1 *Evaluation of Likert scale items*

The possible answers of the Likert scale were encoded with values ranging from 1 for "not interesting at all" to 4 for "very interesting". Whereas the top 20 interests of profile type 1 scored 2.38 ± 0.33, the same selection of interest categories for profile type 2 scored higher with 2.80 ± 0.39. This trend could be found for all n-best selections (see table 5) reaching a maximum difference of 0.7 for the top 5 interests.

Table 5: Mean scores of Likert scale items

	Type 1 $n = 52$		Type 2 $n = 52$	
	M	SD	M	SD
Top 5 interests	2.3808	0.3302	3.0840	0.4361
Top 10 interests	2.3923	0.3318	2.946	0.4166
Top 15 interests	2.3859	0.3309	2.8645	0.4050
Top 20 interests	2.3798	0.3300	2.8021	0.3963

The recommended Wikipedia articles (profile type *recommendations*) have been evaluated with an average score of 2.46 ± 0.33 by the participants.

5.3.2 *Precision*

The precision measures the ratio of relevant recommended items to all recommended items. For calculating the precision we considered items rated as "very interesting" and "interesting" as relevant to the user (true positive) and items rated as "hardly interesting" and "not interesting at all" were considered irrelevant (false positive). Figure 2 depicts the precision curves for different n-best selections of profile type 1 (red curve, dashed) and profile type 2 (blue curve, solid).

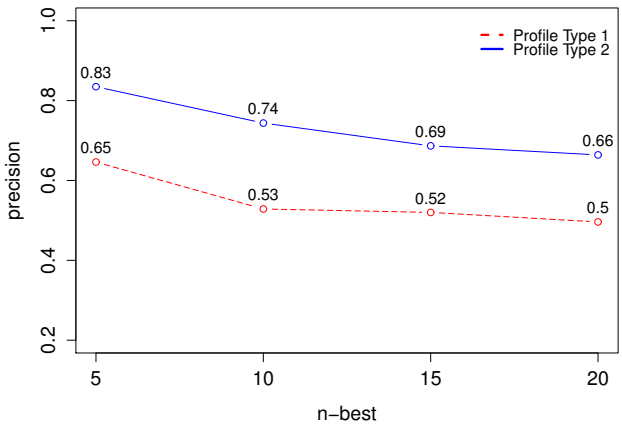

Figure 2: Precision curves profile type 1 and 2

Again profile type 2 (disambiguation heuristics applied) dominates profile type 1 (no disambiguation) in each n-best selection: Regarding the top 5 interests, users indicated a correct assignment for over 80%. For all inferred topics (top 20), at least two thirds are considered relevant by the users. Similarly, two of the top 3 Wikipedia articles recommended in the third step are considered relevant.

5.3.3 *MAP and MRR*

Mean Average Precision and Mean Reciprocal Rank answer the question of how well the interests are ranked at top-k and how early relevant results appear [13]. Again both MAP and MRR scored higher for profile type 2 (0.72 and 0.85) than for profile type 1 (0.50 and 0.68).

5.3.4 *Comparative evaluation with a user's friend*

Profile type 4 was built up with top-10 interest categories of profile type 2 and 10 interest categories where no interest was assumed. The users were asked to evaluate this profile and send a link to a friend of theirs to do the same. The performance of our algorithm and the friend's assessment was compared by the user's evaluation (benchmark). Table 6 shows the confusion matrix comparing the performance of the algorithm introduced in this paper and the user's friends (in brackets).

With a combined success rate of 74 % versus 60 % our approach clearly outperforms the friend in predicting the user's interests. Differences in the above mean values are statistically significant ($p < 0.01$).

Table 6: Performance algo. and friend (in brackets)

	Recommended	Not recommended
Relevant	73 % (55 %)	27 % (55 %)
Not Relevant	24 % (35 %)	76 % (65 %)

5.4 Analysis and Discussion

The results of the user study show that a user's followees are not only a sufficiently broad basis for inferring interest profiles but these interest profiles are also a valid representation of the user's interests. Profile type 2 scored better than profile type 1 in all quality measures calculated. This implies that the disambiguation heuristics have a significant impact on the quality of the generated interest profiles. With over 7 out of 10 items being relevant to the users our approach could achieve state of the art results and performed even better in predicting the users' interests than their friends (thus humans) did.

6. PRODUCTION VERSUS CONSUMPTION BASED PROFILES

We did additional research to answer the introductory question, which also was raised by [17], of how interest profiles based on the user's tweets (production) and followees (consumption) differ. Research on this questions also taps into the long-running debate on consumption vs. production online, where it is often argued that these two actions, particularly with regard to social media and digital content, are inseparable.

6.1 Approach

To provide the basis for a valid comparison of the two different profiles we generated them using the same approach and knowledge base, but extracted the entities from the user's tweets in one case and from his followees in the other. However, we did not conduct a second user study to assess and compare the quality of the two different profile types, but only compare the created profiles.

In the first stage we extracted the entities for the consumption based profile of the user's friend list as described in section 3, whereas for the production based profile the Illinois Wikifier [15], an external tool, was used to extract entities from the user's tweets.
In the subsequent stage, which was the same for both profile types regardless of their input, a spreading activation algorithm was performed on the Wikipedia Bitaxonomy [5] to aggregate the single interest entities to a more abstract and broader interest profile.
Finally, the interest profiles, which represent the user's interests as a list of weighted Wikipedia categories are used to calculate the cosine similarity (as shown in eq. (4)) of the two different interest profiles.

$$sim(X, Y) = cos(\theta) = \frac{X \cdot Y}{||X||\ ||Y||} \tag{4}$$

where X is a vector representing the interest items' weights of the production based profile and Y is the corresponding vector for the consumption based profile. As we use a taxonomy as knowledge base this allows us the comparison of the profiles on different levels of abstractness represented by varying parameters of the spreading activation algorithm (e.g. number of iterations or the decay factor).

6.2 Sample

To calculate the similarity of the two different profile types we selected a random sample of 50 twitter users from the Twitter sample endpoint, which allows to access a small sample of all public statuses and applied a set of selection criteria (e.g. only public accounts, a sufficient number of tweets and friends) on it to get suitable accounts for our experiment only. Apart from that we took the ten participants from the user study described in section 5 that rated the interest profile best and the ten that rated it worst. This leads to a total number of 70 twitter users for which both profile types and their cosine similarity were calculated.

6.3 Results

Compared to the followees of a user it appears to be easier to extract entities from the tweets as results showed that about 2.5 times more entities could be extracted using tweets rather than friends as input. Only about 2 % of the extracted entities (meaning Wikipedia pages) were shared of both sets. Whereas an intersection of 9 % could be found for the first level categories (representing the initially activated nodes). At the first glance these results indicate that the generated profiles and inferred interests do not seem to be too similar.

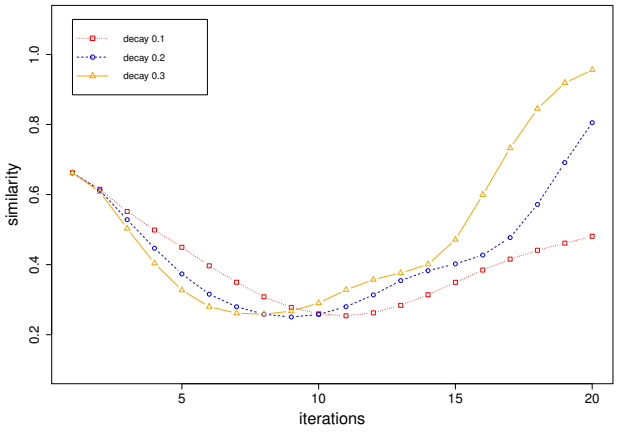

Figure 3: Cosine similarity of production and consumption based profiles for different decay factors

To gain a deeper insight the cosine similarity of the two interest profiles (list of interest categories and their corresponding weight) was calculated with a fixed decay factor (0.1, 0.2 and 0.3) and iterations ranging from 1 to 20. The number of iterations represents the different levels of abstraction, ranging from very concrete (one iteration meaning the initially activated nodes and their corresponding weight) to very abstract (20 iterations upwards in the Wikipedia

category taxonomy). As shown in figure 3 the cosine similarity of the concrete profiles (one iteration) comprising the weighted initially activated nodes is with about 0.66 higher as it would be expected given the small intersection of entities and categories. With an increasing number of iterations the similarity of both profiles is decreasing, whereby the low is reached later if the decay factor is smaller. As expected the similarity is rising for a high number of iterations again, as the categories reached by the spreading activation algorithm now are of a very abstract nature, i.e. the spreading activation accumulates in the top level categories. In general, profiles appear to be more similar on very concrete and abstract levels of the taxonomy.

Figure 4 compares the similarity of the 10 profiles that have been evaluated worst and the 10 that have been evaluated best in the user study that is described in section 5 above.

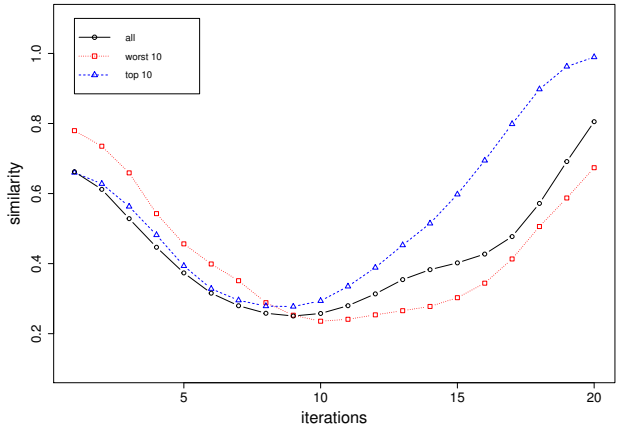

Figure 4: Cosine similarity of production and consumption for 10 best and worst evaluated profiles

The overall trends and the "u"-shape that can be seen in similarity with an increasing number of iterations are the same for both selections. However, the worst evaluated profiles have a higher similarity on a more concrete level (smaller number of iterations) whereas the best evaluated profiles seem to be more similar on an abstract level (higher number of iterations) approaching complete similarity at 20 iterations.

6.4 Analysis and Discussion

In general the cosine similarity of both profiles turns out to follow a "u-shape" along an increasing number of iterations. This means that profiles are more similar on very concrete and abstract levels of the Wikipedia Bitaxonomy (the used knowledge base). While the latter is expectable, since the spreading activation accumulates on the top level categories, the former is an interesting finding. In particular, since the profiles only share 2% of extracted entities. The larger intersection on the first level of categories (9%) is also not surprising, as the categories provide an abstraction of the very specific entities. That is, the number of (available) categories is smaller than the number of (available) entities. However, the high cosine similarity is an interesting finding, which suggests, that even though the intersection of entities and first level categories is low, the weights accumulate in the same categories. Still, the huge amount of categories in which the approaches differ seem to trigger different routes through the category graph to the top levels.

Even though there are differences in the concrete entities that could be extracted from the different sources of input, the information we get based on the consumption and production of the users' twitter accounts are quite similar for certain levels of abstraction. The results therefore support the hypothesis that consumption and production online, particularly with regard to social media and digital content, are inseparable. Nevertheless there is further research needed to support the findings of this work and approach to an explanation for the significant differences in similarity on the different levels of abstractness.

7. CONCLUSION AND FUTURE WORK

In this paper we introduced an approach for inferring semantically meaningful interest profiles from the accounts a user follows on Twitter. Because the followees are the only input used, it is possible to generate interest profiles even for users that posted no tweets. Also, investigating the coverage of followee lists in terms of named entities in Wikipedia revealed that follwees indeed provide sufficient input for creating comprehensive semantic interest profiles. As passive social media use is on the rise, the approach is an important contribution to the development of future social-media-based recommender systems that try to address the cold start problem. By conducting an extensive user study we could show that our approach achieved state of the art (and superhuman) results in predicting a user's interests. A comparison of followee- and tweet-based profiles revealed high similarity on very concrete and abstract levels, suggesting that passive and active use are tightly coupled.

For future work we plan to extend our approach to other social networks such as Facebook (for which "likes" should be semantically equivalent to followees on Twitter). The evaluation showed that the disambiguation heuristics had a significant impact on the profile quality. Hence it appears to be promising use more sophisticated disambiguation and entity linking algorithms in the future and exploit recent advances in that field [21].
Further, we aim to investigate the relation between production and consumption based profiles in more detail. In particular, we are interested in whether a combination of both could improve the performance or if they can mutually benefit from each other.

8. ACKNOWLEDGMENTS

The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601.

9. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized

news recommendations. In *User Modeling, Adaption and Personalization*, pages 1–12. Springer, 2011.

[2] C. G. Akcora, B. Carminati, E. Ferrari, and M. Kantarcioglu. Detecting anomalies in social network data consumption. *Social Network Analysis and Mining*, 4(1):1–16, 2014.

[3] A. Edmunds and A. Morris. The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management*, 20(1):17 – 28, 2000.

[4] S. Faralli, G. Stilo, and P. Velardi. Recommendation of microblog users based on hierarchical interest profiles. *Social Network Analysis and Mining*, 5(1):1–23, 2015.

[5] T. Flati, D. Vannella, T. Pasini, and R. Navigli. Two is bigger (and better) than one: the wikipedia bitaxonomy project. In *Proc. of ACL*, pages 945–955, 2014.

[6] S. Gunelius. Facebook's growing problem - passive users. `http://www.corporate-eye.com/main/facebooks-growing-problem-passive-users/`, 2015.

[7] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth. User interests identification on twitter using a hierarchical knowledge base. In *The Semantic Web: Trends and Challenges*, pages 99–113. Springer, 2014.

[8] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[9] S. LeMole, S. Nurenberg, J. O'Neil, and P. Stuntebeck. Method and system for presenting customized advertising to a user on the world wide web, 1999.

[10] H. Lieberman et al. Letizia: An agent that assists web browsing. *IJCAI (1)*, 1995:924–929, 1995.

[11] K. H. Lim and A. Datta. Interest classification of twitter users using wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, page 22. ACM, 2013.

[12] C. Lu, W. Lam, and Y. Zhang. Twitter user modeling and tweets recommendation based on wikipedia concept graph. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[13] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[14] R. Pochampally and V. Varma. User context as a source of topic retrieval in twitter. In *Workshop on Enriching Information Retrieval (with ACM SIGIR)*, pages 1–3, 2011.

[15] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *ACL*, 2011.

[16] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[17] P. Siehndel and R. Kawase. Twikime!: User profiles that make sense. In *Posters and Demonstrations Track*, ISWC-PD'12, pages 61–64, Aachen, Germany, Germany, 2012. CEUR-WS.org.

[18] L. Tamine-Lechani, M. Boughanem, and N. Zemirli. Inferring the user interests using the search history. In *Workshop on information retrieval, Learning, Knowledge and Adaptatbility (LWA 2006)*, pages 108–110, 2006.

[19] K. Tao, F. Abel, Q. Gao, and G.-J. Houben. Tums: twitter-based user modeling service. In *The Semantic Web: ESWC 2011 Workshops*, pages 269–283. Springer, 2012.

[20] W. Youyou, M. Kosinski, and D. Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[21] S. Zwicklbauer, C. Seifert, and M. Granitzer. Robust and collective entity disambiguation through semantic embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 425–434, 2016.

ABOUT THE AUTHORS:

Christoph Besel received a B.Sc. in Internet Computing from University of Passau, Germany in 2016. He is currently doing his Master's in Web Science and Big Data Analytics at University College London, United Kingdom. His research interests include social media analytics, recommendation systems, applications of data mining and web economics in particular.

Jörg Schlötterer received his bachelor and master degree at the University of Passau, Germany in 2010 respectively 2013. He is currently pursuing his Ph.D. degree in Computer Science at the Professorship of Media Computer Science (University of Passau). His research interests center around information retrieval and connected topics, such as text mining, user profiling and search user interfaces.

Michael Granitzer has been Professor for Media Computer Science at University of Passau since 2012. Before, he was Scientific Director of the Know-Center Graz since 2010 and assistant professor at the Knowledge Management Institute of Graz University of Technology since 2008. In 2011, he was Marie Curie Research Fellow at Mendely Ltd. working on machine learning and information retrieval in academic knowledge bases. His research addresses topics in the field of Knowledge Discovery, Visual Analytics, Information Retrieval, Text Mining and Social Information Systems. He published over 180 mostly peer-reviewed publications and has been scientific coordinator and participant in several EU funded and nationally funded projects.