Undefined 0 (0) 1
IOS Press

The effect of generalization on user interpretation of topical overviews

Alex Olieman [a,b,*], Gleb Satyukov [a,c] Emil de Valk [a]

[a] *Stamkracht, Oostenburgervoorstraat 72, 1018 MR, Amsterdam, The Netherlands*
E-mail: <first_name>@stamkracht.com
[b] *Institute for Logic, Language, and Computation, University of Amsterdam, Amsterdam, The Netherlands*
[c] *Media Technology, Leiden University, Leiden, The Netherlands*

Abstract. The demand for tools that enable interactive exploration of social media streams and other user-generated content has inspired much research in recent years. A common approach in this area starts by extracting information from user contributions, which is subsequently linked to a semantic knowledge base. In this way, entities and concepts that are mentioned in the content are given canonical representations, which serve as the basis to aggregate and compare social media activity over users and over time. While this leads to representations of social media content that can be effectively used behind the scenes of an application, the suitability of these overviews for user interaction has yet to be investigated.

We have conducted an experiment to investigate whether the presentation method that is used to show a topical overview of documents to users has an effect on users' ability to interpret such overview. More specifically, we test for an effect of generalizing topics to a higher level of abstraction on the ease with which users make sense of topical overviews. We found significant effects of this treatment on user accuracy, interpretation diversity, and task duration. Overall, the results indicate that generalization negatively effect users, but we were also able to identify several cases in which generalized overviews were more user-friendly.

Keywords: Topical Overviews, Semantic User Profiling, Concreteness Effect, Human Factors, Wikipedia Categories, Social Media

1. Introduction

One of the key contributions that semantic technologies bring to the search and analysis of social media activity, lies in summarizing the abundance of messages into overviews that can be more easily interpreted. Most existing research in this area has, in one way or another, made use of broad-coverage Knowledge Bases, e.g. as a source of background knowledge to extract topics from social media content–known as Ontology-based Entity Recognition or Entity Linking–as well as to increase the abstraction level of identified keyphrases [1]. These techniques are used primarily to model changes in the frequency with which topics are discussed over time, and to derive interest profiles from a user's activity stream [1].

Several algorithms have been proposed to generate user interest profiles that can be used for clustering, recommendation, and search [2,3,4,5,6]. These use-cases, however, all deal with the utility of interest profiles for further computer processing, and consequently haven't raised many questions about how a user may interact with another's profile. Interactional aspects *do* matter for real-world applications, because users need to inspect and interpret the output of clustering, recommendation, and search functionality.

There is a demand for user-friendly topical overviews of user-generated content in social media, we have found in our work with medium and large enterprises. This demand is, in our experience, not limited to user interest profiles, but also includes overviews of group

*Corresponding author. E-mail: alex@stamkracht.com.

activity streams. The observation that *"the main challenge in browsing and visualisation of high-volume stream media is in providing a suitably aggregated, high-level overview,"* [1] leads us to questioning how high-level an overview should be. Creating overviews necessarily involves abstraction or generalization, but which degree of abstractness is suitable for a given task?

Earlier work on tagging in social media has shown that abstract tags are, on average, more frequently used than concrete tags [7]. Moreover, when users judge their own interest profiles they rate general entities as more relevant than specific entities [8]. Finally, profile sparsity can be reduced by generalizing topics, for instance to increase recall in search, to generate a larger number of recommendations, or to cluster profiles that are only related by their broader topics [3].

Other disciplines, however, provide compelling evidence which suggests that more specific or concrete topics are preferable when topical overviews will be presented to users. It has been repeatedly observed *"that concrete nouns are processed faster and more accurately than abstract nouns in a variety of cognitive tasks."* [9] This is referred to as the *concreteness effect*, for which psychological literature offers two competing–but mutually non-exclusive–explanations. The first, dual-coding theory, claims that abstract words are only processed verbally, whereas concrete words are additionally processed by an image-based system which significantly assists the working memory [9,10,11]. Context-availability theory, on the other hand, claims that concrete words allow easier access to their semantic context, which leads to faster processing without a dependence on non-verbal brain regions [9,10,11].

Neuroimaging studies have found evidence for both proposed causes of the concreteness effect, on a range of tasks [9,10,12]. Even though questions about the suitable level of generalization for topical overviews are not reducible to a discussion about the concreteness effect, it seems likely that the concreteness effect plays a role when users interpret individual topics and form an overall impression from them.

We therefore conducted an experiment to test whether the incorporation of generalized topics into topical overviews has an effect on users who were given a task in which they needed to interpret the overviews.

This paper is structured as follows: First, we provide a survey of existing research about topical overviews of documents, and particularly about the more specific case of generating user interest profiles from social media streams. Section 3 describes our approach for generating topical overviews, for a flat presentation as well as a generalized presentation method. In Section 4 we discuss the experimental task and procedure that were used to perform a user study. The subsequent section summarizes the results of the user study, and is followed by a discussion and implications for future work.

2. Related work

This survey of related literature is focused on studies that use Wikipedia articles and categories as canonical representations of topics, which we ourselves do to generate and generalize overviews. Several of the cited studies use DBpedia, rather than–or in combination with–directly using data from Wikipedia, but we refer to Wikipedia throughout for the sake of consistency.

2.1. Topical overview of documents

Early approaches for enriching documents with topics that are represented by Wikipedia articles and categories have shown that this could increase classification performance [13,14]. Wikipedia categories can also be used to identify topics that are common to a set of documents. The motivation here is to enable users to gain a topical overview of the collected readings of other users [15]. This technique is generic enough to be applied to other kinds of document sets, e.g. search results or social media conversations. Syed et al. found that spreading activation with two pulses led to the best results for this task [15].

A similar technique can be used to automatically generate labels for document clusters that have been found by unsupervised methods. Even without using the Wikipedia category graph, the labels generated by the category-based method were of a higher quality than labels extracted from the documents [16].

In an Information Retrieval context, topical overviews of individual search results can give a quick impression of what a document is about. This is complementary to the now-common "snippets," which show only the relation between the query and the document. In [17], the generation of "compact representations" of documents (i.e. topical summaries) is investigated for the chemistry domain. They create topical summaries from Wikipedia categories, as well as from an ontology of chemical entities, and let a group

of domain experts evaluate them. The category-based method, on average, performs better than the domain ontology-based method. A notable difference is that for Wikipedia broader categories quickly became less relevant, whereas for the domain ontology this decline was only significant after traversing the ontology more than three levels upwards [17].

2.2. User interest profiles

Szomszor et al. have investigated linking tags from a range of social networking sites to Wikipedia categories, as a means to unify distributed user profiles [2]. Their approach consists of harvesting user tagging activity, and string matching these tags to Wikipedia articles (without disambiguation). Subsequently, they collect the categories in which these articles are included, but only those whose names closely match the source tags, or when an article is a single category. A single user's profile is formed by calculating the sum of tag frequencies per category, and discarding categories with below-average frequency [2].

Michelson and Macskassy generate interest profiles for Twitter users by discovering mentioned entities in tweets, disambiguating them, and linking them to corresponding Wikipedia articles [3]. Subsequently, a tree is constructed for each found entity, representing its category memberships, and their more general (i.e. parent, broader) categories up to 5 levels. To construct the user profile, entity and category occurrence are counted, and categories are ranked by score: $Score(c) = Freq(c) * b^{-d}$, where d is the depth of the category in the tree, and b is a constant branching factor. The user profile is finally formed by selecting the top-k categories [3].

Kapanipathi et al. [4] similarly view broader categories as potentially relevant for user interest profiles. They have investigated several spreading activation functions as means to propagate interest scores, based on frequency, from linked articles (*Primitive Interests* in [4]) to the more general categories that contain them (*Hierarchical Interests* in [4]). This approach was evaluated in a user study with 37 participants who judged their own profiles, which consisted of the top-50 hierarchical interests. The spreading activation function that performed best takes the order in which categories are listed at the bottom of a Wikipedia article into account, and boosts categories which have multiple activated sub-categories [4].

Abel et al. address the issue of limited context in microblog posts, by following links from tweets to news articles [5]. They extract entities from the tweets themselves, but demonstrate that linked news articles yield a larger number of entities, and a greater diversity of types of entities. This leads to user profiles that may be more representative of a Twitter user's interests. In [18] this approach is extended with a GUI which displays the most frequently occurring entities for a selected profile, as well as an overview of more general topics that are associated with the discovered entities.

Another extension of the entity linking approach to generating user interest profiles takes the temporal nature of interests into account. Orlandi et al. model this temporality by applying an exponential decay function to interest weights [6]. They found, by letting 21 participants judge their own profiles, that a mean lifetime of 360 days generated profiles that were rated higher by participants than those generated with a shorter mean lifetime of 120 days. This difference, however, could not be confirmed as statistically significant. Participants did rate profiles based directly on discovered entities significantly higher than profiles based on these entities' categories. Orlandi et al. remark that *"according to the results, we think that mixed approaches adopting both categories and [entities] for user profiling can be highly beneficial and need to be investigated."* [6] Such mixed approaches could possibly address the issue that entities are often overly specific, whereas categories can be too broad to be relevant.

Shen et al. incorporate user interest profiles into their approach for collective entity linking on tweets [19]. They generate initial interest profiles from the entities that were detected in a user's tweets (pre-disambiguation), and represent all candidate entities together with their semantic relatedness in a graph. In this way, the limited context in microblogs can be partially overcome by incorporating the topical coherence between candidate entities mentioned in a new tweet and in previous tweets into the disambiguation algorithm.

3. Generation of topical overviews

This section describes our approaches for creating topical overviews from arbitrary sets of source documents. The first approach is fairly simple, and results in a flat list of ranked topics. The second approach builds on the first, and outputs a two-level nested list.

3.1. Flat presentation

This baseline approach consists of only two steps. First take the set of source documents, find the substrings that mention entities and concepts, disambiguate them, and link them to DBpedia. We use DBpedia Spotlight [20] to perform this Entity Linking (EL) step. Subsequently count how many source documents link to each of the entities, and sort the entities from highest to lowest frequency. Ties are resolved alphabetically on the entity labels.

Apart from the EL method, this approach is equivalent with "entity-based profiles" in [5]. It is also similar to "resource-based profiling" in [6], but doesn't use a time decay function.

3.2. Generalized presentation

For this more elaborate presentation method we follow several suggestions that were found in related work. The visualization should support different levels of granularity [1] (i.e. specific–general). It should feature a combination of entities and categories, and similar topics should be clustered [6]. Lastly, it should be compact, by collapsing the overviews into a few representative phrases [3].

Let A be the set of Wikipedia category URIs, and B be the set of Wikipedia article URIs ($A \cup B = \emptyset$). Let $E \subset B$ be the set of entities that the source documents link to. For each $e \in E$, find parent categories $C_e \subset A$, up to m levels outward. In DBpedia the links from entities to categories are represented as `dcterms:subject`, and from categories to their parents as `skos:broader`. Together, the (grand)parent categories form the set

$$C = \bigcup_{e \in e} C_e \tag{1}$$

The traversal which finds C_e with $m = 3$, can be done with the SPARQL query in Code 1. This example can be run on the `http://dbpedia.org/snorql` interface.

In order to find representative categories for clusters of entities, let $D = (d_{ce} : c \in C, e \in E)$ be a $|C| \times |E|$ sparse matrix which contains category–entity distances d_{ce}, where d_{ce} is the length of the `skos:broader` path between e and c, and $0 \leq d_{ce} \leq m$. To continue the SPARQL example, where $e =$ `:View_of_Delft`, we find (among others):

Code 1: Example query to find C_e

```
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT DISTINCT ?c0 ?c1 ?c2 WHERE {
  VALUES ?e { :View_of_Delft } .
  {
    ?e dcterms:subject ?c0.
  } UNION {
    ?e dcterms:subject/skos:broader ?c1.
    FILTER NOT EXISTS {
      ?e dcterms:subject ?c1.
    }
  } UNION {
    ?e dcterms:subject/skos:broader
      /skos:broader ?c2.
    FILTER NOT EXISTS {
      ?e dcterms:subject ?c2.
    } .
    FILTER NOT EXISTS {
      ?e dcterms:subject/skos:broader ?c2.
    }
  }
}
```

$d_{:Category:Delft,e} = 0$, and $d_{:Category:Paintings,e} = 2$.

D is initialized with $null$ values, and has the index set C for rows and E for columns. Similarly to Kapanipathi et al.'s "Intersect Booster" [4], we want to identify categories that have many transitive inlinks from entities, with the shortest possible path distances.

To order C by suitability to represent a cluster of entities, we define:

$$ParentRank(c, D) = \frac{\gamma + \sum_{e \in E} d_{ce}}{Coverage(c, D)} \tag{2}$$

where

$$Coverage(c, D) = |(d_{ce} : d_{ce} \neq null, e \in E)|$$

$$\gamma = \frac{\kappa}{Coverage(c, D)}$$

A lower $ParentRank$ value, following its definition, indicates that a category is representative of a larger cluster of entities. The penalty γ is applied to reduce the likelihood of ties, and κ is a constant for which higher values favor the number of inlinks over the row-wise sum of d_c. In this paper we use $\kappa = 1$. Because every Wikipedia article is directly contained in at least one category, it follows that $\forall_{e \in E} \exists_c d_{ce} = 0$.

Algorithm 1 is used to form clusters of entities that are be paired with a category that should be represen-

Algorithm 1. Cluster entities under categories.
Require: Φ, an iterator over sorted rows of D
1: $toAssign \leftarrow E$
2: $assigned \leftarrow list()$
3: **for each** $c, d_c \in \Phi$ **do**
4: $\quad inBoth \leftarrow toAssign \cap indices(d_c)$
5: $\quad$ **if** $|inBoth| > 1$ **then**
6: $\quad\quad assigned.append(< c, indices(d_c) >)$
7: $\quad\quad toAssign \leftarrow toAssign \setminus inBoth$
8: $\quad\quad$ **if** $toAssign = \emptyset$ **then**
9: $\quad\quad\quad$ **break**
10: $\quad\quad$ **end if**
11: $\quad$ **end if**
12: **end for**
13: $orphans \leftarrow toAssign$
14: **return** $assigned, orphans$

tative of the members of the cluster. Finally, the resulting clusters are sorted by the sum of their source document counts, from high to low. The entities that aren't members of any cluster are added at the bottom. We present this generalized topic overview to users as a nested list, with category labels on the top level, and the underlying entities on the second level.

4. User study

In this section we discuss the randomized experiment that was performed to investigate our main research question:

Does the addition of generalized topics to overviews help users to perform a task in which they need to make sense of the underlying user-generated content?

We formulate hypotheses for three distinct variables of interest:

A) The ability of users to interpret topical overviews in correspondence with the source documents:
   - H_0: The inclusion of generalized topics has no effect on the accuracy with which users interpret overviews.
   - H_1: The inclusion of generalized topics has a negative effect on user accuracy.
   - H_2: The inclusion of generalized topics has a positive effect on user accuracy.

B) The amount of diversity in users' interpretation of the same topical overview:
   - H_0: The inclusion of generalized topics has no effect on interpretation diversity.
   - H_1: The inclusion of generalized topics increases interpretation diversity.
   - H_2: The inclusion of generalized topics decreases interpretation diversity.

C) The amount of time that users take to interpret a topical overview:
   - H_0: The inclusion of generalized topics has no effect on task duration.
   - H_1: The inclusion of generalized topics increases task duration.

Regarding user accuracy and interpretation diversity we consider that the effect of generalization could go both ways. If the concreteness effect plays a significant role for users who interpret the overviews, we would expect that user accuracy decreases and interpretation diversity increases. However, if users prefer general topics in these overviews, as they do in their own profiles, or benefit from the nested presentation, we would expect the opposite effect. We expect an increase in task duration, in line with the concreteness effect. Even though the generalized overviews are more compact on first sight, they offer an additional interaction mechanism (i.e. expand/collapse category nodes), and contain more information than the flat overviews.

The user study is designed as a remote experiment such that participants are free to decide where and when to take part.

4.1. Participants

Sixty-four participants, aged 20-61 (μ=31.8; SD=9.7), of which 26 female, completed our experimental procedure. A large majority of participants was highly educated. They were recruited via convenience sampling by an invitation that was spread through online social networks, originating from the authors and at least eight of their co-workers, and was shared from there on.

4.2. Task

The choice to conduct a task-directed experiment was motivated by the observation that there has been a lack of task-based evaluation in the literature on topical overviews of social media streams [1]. In order to measure the variables of interest, we implemented a manual classification task for our remote user study. This task is framed as an expert finding scenario in which the participants are looking for journalists with var-

ious specializations. The hypothetical search system, however, does not allow for a more specific query than "journalist." This is used as an excuse to let the participants evaluate a sequence of user profiles, which they need to interpret in order to classify them into one of five classes: Art, Finance, Sports, Technology, Travel, or into an "Other" class. We argue that this task is an acceptable compromise between practical considerations for a controlled experiment and the representativeness for realistic user tasks.

In the experimental condition participants are asked to classify generalized user profiles. The profiles in question are presented as a nested list that is two levels deep and is initially in a collapsed state. This type of widget is commonly known as an "accordion". The first level of the accordion consists of the generalized topics, and clicking on one of these topics reveals the sub-list of more specific topics.

In the control condition participants are presented with regular user profiles–without added categories–presented as a flat list of the specific topics that were extracted from the source documents.

In both conditions the profiles are displayed in the middle of the screen and order of topics is as described in the algorithm in Section 3.2.

4.3. Fictional profiles

We have created 18 profiles of fictional journalists, which are enumerated in Table 1. The profile IDs indicate the sequence in which participants saw each profile. This sequence was randomized once; before we opened the study to participants. L_C and L_E are the profile lengths in the control and experimental condition, expressed as the number of topics that are visible when the profile is loaded.

The source documents for the profiles were selected by searching within the sites of Dutch news publishers that are geared towards the given classes. No suitable news publisher could be found for the Travel class, therefore travel blogs were used instead. These sites were searched with terms that indicate recognizable sub-topics within each class. For each profile, 3-5 source documents were selected (depending on length), preferably by the same author. We added fictional names to the profiles to reinforce the idea that these were profiles of different users, without giving away hints about the correct class.

Table 1: Fictional user profiles

ID	Category	Description	L_C	L_E
1	Other	Local News	20	6
2	Finance	Stock Exchanges	28	11
3	Sport	Dressage	13	6
4	Sport	Cycle Sport	5	2
5	Art	Museums	22	9
6	Other	Fashion	23	7
7	Technology	Home Automation	94	32
8	Technology	Online Privacy	49	14
9	Art	Dutch School (painting)	11	4
10	Art	Art Nouveau	70	25
11	Finance	Housing Markets	9	5
12	Other	Animals	21	11
13	Travel	South-American Travel	25	6
14	Travel	Asian Travel	35	13
15	Sport	Judo	7	3
16	Technology	Smartphones	63	25
17	Travel	Eastern-European Travel	35	14
18	Finance	Foreign Exchange Market	13	7

4.4. Experimental procedure

For the evaluation environment we set up a typical (of the modern day) web-application based on the MEAN stack[1]. It consisted of a sequence of 6 screens:

1. Welcome page
2. Registration page
3. Introduction page
4. Competence test
5. Experimental task
6. Reflection page

On the welcoming screen participants were invited to enroll and were given a brief outline of the study. Next, on the registration page, participants entered their demographic attributes. They were blindly assigned a random condition once they proceeded to a detailed explanation of the experiment.

Subsequently, in order to test for familiarity with the "drag & drop" interaction mechanism, participants were required to drag images of a few highly recogniz-

[1] https://en.wikipedia.org/wiki/MEAN_(software_bundle)

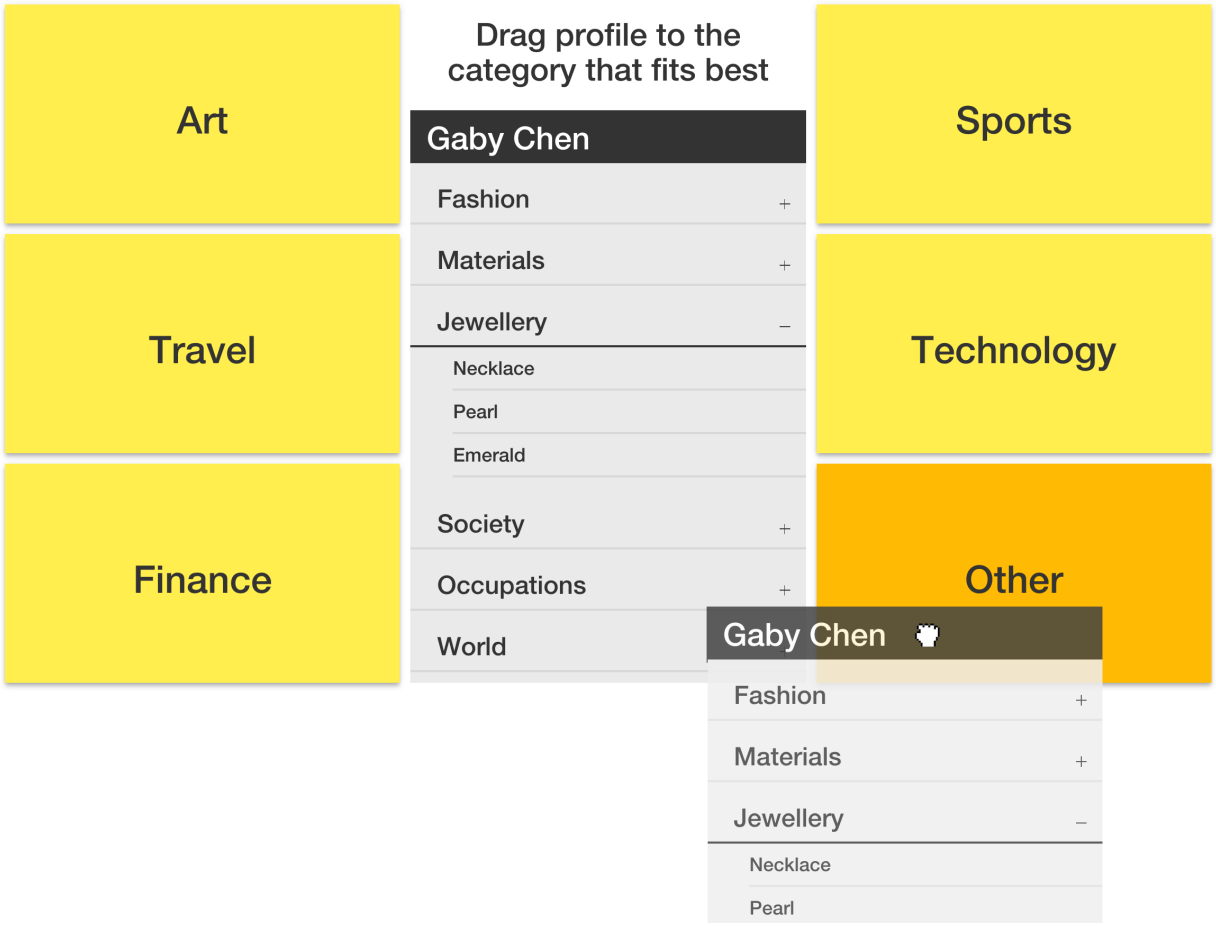

Fig. 1. Screenshot of the user study GUI, in which a generalized profile is being classified.

able objects (e.g. tree, cat, apple) into corresponding classes (e.g. trees, cats, fruit). As soon as a participant passed this competence test, he/she was redirected to the experimental task.

During the task participants were presented with the profiles described in 4.3. The classes were displayed as relatively large rectangles positioned on both sides of the profile. Participants were shown one profile at a time, which they dragged and dropped into the class that best matched their interpretation of the profile. During this task we recorded the duration between the profile being displayed and when it was classified, as well as into which class the profile was dropped.

After the task was completed, participants landed on a page where they were asked to answer some questions about the task. They were asked to indicate how certain they were of their classifications for each of the categories (1=*very unsure* to 5=*very sure*). They were also asked to indicate their (dis)agreement with three statements (1=*completely disagree* to 5=*completely agree*): *the names of the topics were clear*, *the topics were arranged usefully*, and *the topics gave a clear picture of what kind of journalist the profile belonged to*. Finally, participants were invited to add any remarks about the study.

Various technical considerations were taken into account to ensure that the experience would be consistent for all participants. The type of device and screen size were checked before prospective participants were allowed to enroll, in order to minimize variation of interaction modalities, and to control the proportion of screen area in which the task was displayed. More specifically: the use of mobile devices was not permitted, and the size of the inner browser window needed to be at least 1200 pixels wide and 720 pixels tall.

It is also worth noting that because the study was carried out in the Netherlands, it was decided to display the experimental materials in Dutch. This lan-

guage was preferred over English because of its high percentage of native speakers, so as to reduce the risk of language comprehension as a confounder.

5. Results

5.1. User accuracy

Participants who were presented with generalized profiles were on average less accurate (0.71 ± 0.05) than participants who saw the flat profiles (0.77 ± 0.05). This difference is significant with $p < 0.05$, as determined by a Mann–Whitney U test.

The difference in accuracy is unevenly distributed over the 18 profiles. By calculating per-profile accuracy as the proportion of participants that classified the profile correctly, the difference between the conditions could be analyzed in more depth. Figure 2 shows profile accuracy of the experimental group minus that of the control group.

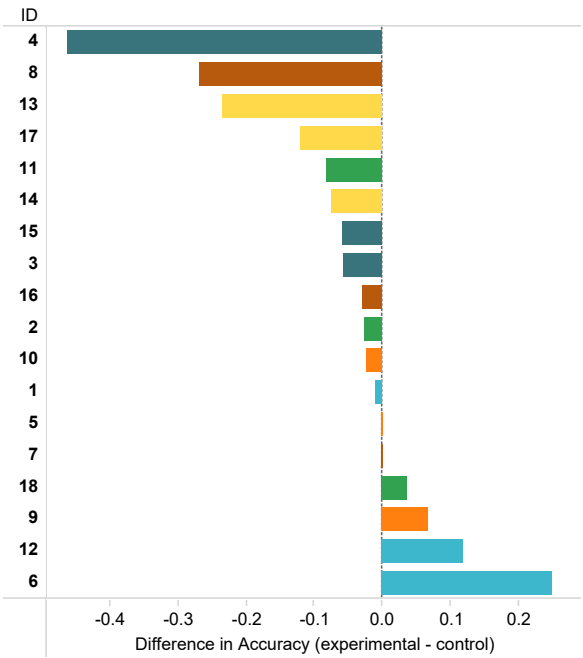

Fig. 2. Per-profile difference in accuracy between the conditions. The bottom rows show which profiles were more accurately classified by the experimental group. Color codes for actual class.

5.2. Interpretation diversity

In the context of this study, interpretation diversity is operationalized as: the uniformness of the distribution of participant misses over the 5 incorrect classes, for a given profile. We took, for each profile and per condition, the proportional abundance of the incorrect guesses, and compute Shannon entropy H' of this distribution as a diversity index. A pairwise comparison of the resulting diversity values (see Table 2) reveals that 13 out of 18 profiles are more diversely interpreted in the experimental condition. This difference is found to be significant with $p < 0.05$ by a Wilcoxon signed-rank test.

The reflection of participants on the given task indicates that the experimental group found it more difficult to interpret the profiles. Figure 3 summarizes participant (dis)agreement with the statements from Section 4.4. The topic names were judged to be clear by 61% of the control group, versus 48% of the experimental group. For 35% of the control group the profiles gave a clear picture of the represented persons, whereas this was only the case for 18% of the experimental group. The nested presentation of profiles, however, was rated as useful by 24% of participants, while the majority of participants (61%) who received the flat profiles disagreed that the topics were arranged usefully.

Table 2: Diversity and duration per profile

ID	H'_C	H'_E	$\Delta H'$	μ_C	$\Delta\mu$	$\tilde{p}$
1	0.00	0.89	+0.89	26.03s	+12.58s	0.035
2	0.00	0.64	+0.64	8.02s	+2.53s	1
3	0.00	0.69	+0.69	7.39s	+1.25s	1
4	0.92	0.47	-0.45	6.47s	+5.66s	0.008
5	0.00	0.00	+0.00	8.75s	-2.59s	0.129
6	0.69	0.79	+0.10	14.48s	-2.84s	1
7	0.00	0.00	+0.00	6.55s	-1.07s	1
8	0.00	0.47	+0.47	5.22s	+6.09s	0.001
9	0.00	0.69	+0.69	7.43s	-3.73s	0.000
10	0.69	0.95	+0.26	7.09s	-0.66s	1
11	0.50	1.07	+0.57	8.48s	-1.66s	1
12	0.00	0.86	+0.86	8.80s	+4.93s	1
13	0.68	0.71	+0.03	8.64s	+2.15s	1
14	0.69	0.30	-0.39	7.00s	+2.28s	1
15	0.00	0.64	+0.64	4.69s	-0.89s	0.408
16	0.00	0.69	+0.69	4.19s	+0.19s	1
17	0.23	0.57	+0.33	9.22s	+2.97s	1
18	0.64	0.00	-0.64	4.24s	-0.11s	1

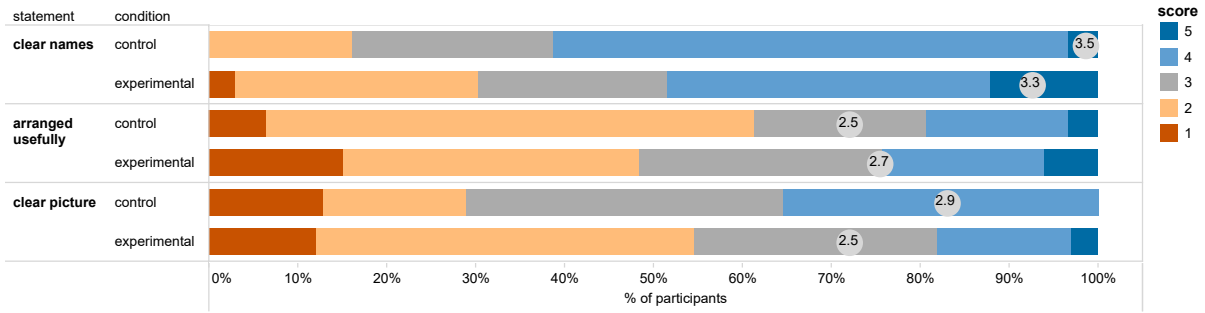

Fig. 3. Participant responses to statements (see Section 4.4); 1=*completely disagree*, 5=*completely agree*.

5.3. Task duration

Participants in the experimental group, on average, required more time to decide which class was the best match for a given profile. Observations with values greater than Upper quartile + 1.5 Interquartile range (per profile*condition) were considered to be outliers; 7% of measured durations were removed this way. Micro-averaged per condition, participants in the control group took 8.57±0.60 s, versus 10.26±0.92 s in the experimental group. This difference could not be confirmed as significant with a Mann–Whitney U test, at $\alpha = 0.05$.

As with accuracy, the differences in duration between conditions are not evenly distributed across the profiles. This was followed-up by testing for significant difference between conditions for each of the profiles. We paired the classification durations per profile, and report on Holm–Bonferroni corrected significance $\tilde{p}$ of independent Mann–Whitney U tests in Table 2. Participants in the experimental group required significantly more time to classify 3 out of 18 profiles, however they classified one profile significantly faster than the control group.

6. Discussion

The results of our experiment suggest that the concreteness effect plays a measurable role in the ability of users to interpret topical overviews. We have found support for the hypothesis that the inclusion of generalized topics has a negative effect on user accuracy. The per-profile analysis of differences in classification accuracy between the groups shows that there are several profiles which contra-indicate this overall effect.

Further qualitative analysis of the difference in presentation of these profiles has led to two interesting observations. Of the top-3 profiles for which the negative effect of generalization was greatest, profiles 4 and 13 featured the topics that were most indicative of the actual class at the bottom of the profile, because they could not be clustered. Of the top-3 profiles that showed a beneficial effect of generalization, profiles 6 and 9 featured neat clusters that clearly indicated the salient topics of the source documents, whereas misleading topics were hidden at the bottom of the profile.

The findings regarding interpretation diversity show that generalized profiles lead to significantly more diverse misinterpretations. In the control group there were 10 profiles for which only one of the possible kinds of error was made. In the experimental group there were just 3 profiles for which only one incorrect class was chosen by the participants. These findings, in combination with participants' reflections on the task, suggest that participants experience a greater sense of ambiguity when interpreting generalized profiles. Furthermore, the per-profile results for accuracy and interpretation diversity are not correlated, which indicates that these variables measure distinct factors.

We were not able to establish a significant difference in the overall task duration between the groups, even though the differences in means and their 95% confidence intervals indicate that there is a fairly large difference. The lack of a significant test outcome may well be caused by a large difference in the variances between the groups. Of the 4 significant per-profile differences, the signs (i.e. +/-) for 3 of the profiles correspond to the signs of the difference in accuracy. This could indicate that accuracy and duration are dependent measures of interpretation difficulty.

We conclude this discussion with two challenges that we encountered. Several participants reported that the introductory text was too long. Moreover, one participant remarked that she only found out that she could expand the general topics near the end of the

task. It is quite likely that other participants in the experimental group, did not notice this possibility at all. While an essential requirement is that participants fully comprehend the task that they are about to perform–having an extended introductory text may lead participants to skip to the task prematurely.

With respect to the manual classification task, one of the participants remarked that "binary classification does not match the reality of people and their interests" [paraphrased]. In retrospect, we could have asked participants to assign a relevance score to each of the categories for each profile (e.g. on a scale of 0-5, as in [6]), rather than asking for binary judgments. This alternative approach would be a more realistic illustration of any potential confusion/ambiguity that participants might experience in their interpretation of a profile. Such a relevance score could also be combined with a fixed stimulus duration, as an alternative to letting participants decide how long they would take.

7. Future work

Our findings suggest that users benefit from being able to see specific topics that are not hidden behind more general topics. When the clustering algorithm does a good job, however, users do seem to find the generalized profiles easier to interpret. We would like to investigate, in future research, whether a new presentation method can benefit from both strengths. Our current suggestion for this novel presentation method is to switch the foreground and background of the generalized presentation method.

Such a clustered overview would show the k most frequently occurring topics within each cluster, and use their parent topic label merely to indicate what other kinds of topics can be found in the cluster. A mock-up of this presentation method is displayed in Figure 4, which shows the same profile as Figure 1. The indented lines below the top-k topics would, as we imagine, reveal all topics in the cluster when they are clicked.

Fashion
Knitting
Catwalk
and 6 more topics in *Fashion*
Metal

Glass
Wool
and 2 more topics in *Materials*
Ode
Futurism
Champagne
and 1 more topic in *Miscellaneous*
Necklace
Pearl
Emerald
in category *Jewellery*
...

Fig. 4. Example of clustered profile presentation.

In addition to introducing a clustered overview of user interests it would be worthwhile to incorporate different presentation modes–as discussed in this paper–into longitudinal *in-situ* evaluation on an social platform. Such an integration would allow users to switch between different types of presentation modes depending on the task at hand. By recording users' choice of presentation mode, we would be able to investigate possible correlation between specific type of overview and different types of activities.

Another advantage of an in-situ evaluation is that the (choice of) presentation mode could be informed by the level of relevant domain knowledge of the user. As an example consider a user who is looking for someone with knowledge in a domain that he or she is very familiar with. Returning an overview that describes user profiles in rather abstract terms may not be satisfactory. On the other hand, a user who ventures into an area where they have little relevant knowledge could be puzzled by a list of domain-specific entities or concepts.

References

[1] Kalina Bontcheva and Dominic Rout. Making sense of social media streams through semantics: A survey. *Semantic Web*, 5(5):373–403, 2014.

[2] Martin Szomszor, Harith Alani, Ivan Cantador, Kieron O'Hara, and Nigel Shadbolt. Semantic Modelling of User Interests Based on Cross-Folksonomy Analysis. *The Semantic Web - ISWC 2008, 7th International Semantic Web Conference*, pages 632–648, 2008.

[3] Matthew Michelson and Sofus a Macskassy. Discovering users' topics of interest on twitter: a first look. *AND '10: Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 73–80, 2010.

[4] Pavan Kapanipathi, Prateek Jain, Chitra Venkataramani, and Amit Sheth. User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In Valentina Presutti, Claudia D'Amato, Fabien Gandon, Mathieu D'Aquin, Steffen Staab, and Anna Tordai, editors, *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*, pages 99–113. Springer International Publishing, 2014.

[5] Fabian Abel, Qi Gao, Geert J. Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6643 LNCS(PART 2):375–389, 2011.

[6] Fabrizio Orlandi, John Breslin, and Alexandre Passant. Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th International Conference on Semantic Systems - I-SEMANTICS '12*, page 41, 2012.

[7] Dominik Benz, Christian Koerner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. One Tag to Bind Them All: Measuring Term Abstractness in Social Metadata. *Semantic Web: Research and Applications, Pt Ii*, 6644:360–374, 2011.

[8] Fabrizio Orlandi, Pavan Kapanipathi, Amit Sheth, and Alexandre Passant. Characterising concepts of interest leveraging Linked Data and the Social Web. *Proceedings - 2013 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2013*, 1(i):519–526, 2013.

[9] F Jessen, R Heun, M Erb, D O Granath, U Klose, a Papassotiropoulos, and W Grodd. The concreteness effect: evidence for dual coding and context availability. *Brain and language*, 74(1):103–112, 2000.

[10] K. Fliessbach, S. Weis, P. Klaver, C. E. Elger, and B. Weber. The effect of word concreteness on recognition memory. *NeuroImage*, 32(3):1413–1421, 2006.

[11] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–11, 2014.

[12] J Oliveira, M V Perea, V Ladera, and P Gamito. The roles of word concreteness and cognitive load on interhemispheric processes of recognition. *Laterality*, 18(2):203–15, 2013.

[13] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E K Park, and Xiaohua Zhou. Exploiting Wikipedia as external knowledge for document clustering. *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (2009)*, 28(1):389, 2009.

[14] Peter Schönhofen. Identifying document topics using the wikipedia category network. *Web Intelligence and Agent Systems*, 7(2):195–207, 2009.

[15] Zareen Saba Syed, Tim Finin, and Anupam Joshi. Wikipedia as an Ontology for Describing Documents. *Artificial Intelligence*, pages 136–144, 2008.

[16] David Carmel, Haggai Roitman, and Naama Zwerdling. Enhancing cluster labeling using wikipedia. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, (January 2016):139–146, 2009.

[17] B. Köhncke and W.-T. Balke. Using wikipedia categories for compact representations of chemical documents. In *CIKM 2010*, pages 1809–1812, 2010.

[18] Ke Tao, Fabian Abel, Qi Gao, and Geert-Jan Houben. TUMS: Twitter-Based User Modeling Service. In *ESWC 2011 Workshops*, volume 7117. 2012.

[19] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling. In *KDD '13*, 2013.

[20] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proc. of I-Semantics 2013*, pages 3–6, Austria, Graz, 2013.