# Deriving Knowledge Profiles from Twitter

Claudia Hauff and Geert-Jan Houben

WIS, Delft University of Technology, Delft, the Netherlands
{c.hauff,g.j.p.m.houben}@tudelft.nl

**Abstract.** E-learning systems often include a personalization component, which adapts the learning content to the learner's particular needs. One obstacle to personalization is the question of how to obtain a learner profile for a learner who just starts using an E-learning system without overwhelming her with questions or unsuitable learning material. One possible solution to this problem lies in the social Web. If a learner is active on the social Web, a considerable amount of information about her is already available. Depending on the social Web service(s) the learner uses, her tweets, photos, bookmarks, etc. are publicly accessible. We investigate if it is feasible to exploit the social Web, more specifically the social Web service Twitter, to infer a learner's knowledge profile in order to overcome the "cold-start" problem in E-learning systems.

## 1 Introduction

Platforms that facilitate E-learning have become increasingly prevalent in recent years. Due to the ubiquitous nature of the Internet, learning online at your own pace and at your own time has never been easier. E-learning systems may include a personalization and adaptation component, which adapts the learning content to the learner's needs and capabilities. Adapting an E-learning system based on a learner's profile can increase learner satisfaction and decrease learner frustration. For example, learning units the learner is already knowledgeable about can be automatically removed by the system, while content the learner is unfamiliar with can be covered in greater depth. One obstacle to personalization is the question of how to obtain a learner profile for a learner who is new to an E-learning system. Although it is possible for the system to derive the learner's knowledge profile over time or by posing a series of test questions, the learner may be unwilling to spend a lot of effort on this procedure. By the time an adequate knowledge profile of the learner has been aggregated, the learner might have already given up on the system.

One potential solution to this problem lies in the social Web. The rise of the social Web has made people not merely consumers of the Web, but active contributors of content. Widely adopted social Web services, such as Twitter[1], Flickr[2] and Delicious[3], are frequented by millions of active users who add, comment or vote

[1] http://www.twitter.com/

[2] http://www.flickr.com/

[3] http://www.delicious.com/

on content. If a learner is active on the social Web, a considerable amount of information about her is available on the Web. Depending on the social Web service(s) the learner uses, her blog entries, tweets, bookmarks, etc. are publicly accessible. For the future, we foresee the following scenario: a learner who uses an E-learning system for the first time, is asked by the system to list her handles of the social Web services she is active on. Then, based on the learner's "online persona", aggregated from the social Web, the system can automatically infer a basic profile of the learner's knowledge in the desired domain. It needs to be stressed, that we envision this approach to be mostly applicable in "cold-start" situations, that is, when the system has no other information available about the learner.

The assumption behind this vision is of course, that it is possible to extract a basic learner's knowledge profile from the social Web. In this paper we investigate this very assumption. More specifically, we investigate if it is feasible to infer a learner's knowledge profile from her activities on the micro-blogging platform Twitter. Our motivation for exploiting Twitter is based on the fact that it is a highly popular platform, used by millions of people[4]. Moreover, most Twitter users make their microblog posts (tweets) publicly accessible, and thus there will be few privacy concerns. The E-learning system only needs to query the learner for her handle on Twitter, no further information (that is, no login information) is required. As many people use Twitter throughout the day, we postulate that among all the posts a user publishes, at least some of them will be pertinent to the user's work and study. Examples of tweets that we envision to be useful are:

- *The Microwave Toolbox for Scilab v0.3 Available for Scilab and Scicoslab.*
- *In algebra 4*
- *Confident I did perfect on my algebra 2 test.*

These tweets on the one hand allow us to infer what the learner is currently learning (*In algebra 4*), but they also allow us to build to some extent a knowledge profile of the learner; the tweet *Confident I did perfect on my algebra 2 test.* implies a high level of knowledge in this particular study area according to the learner's self-assessment.

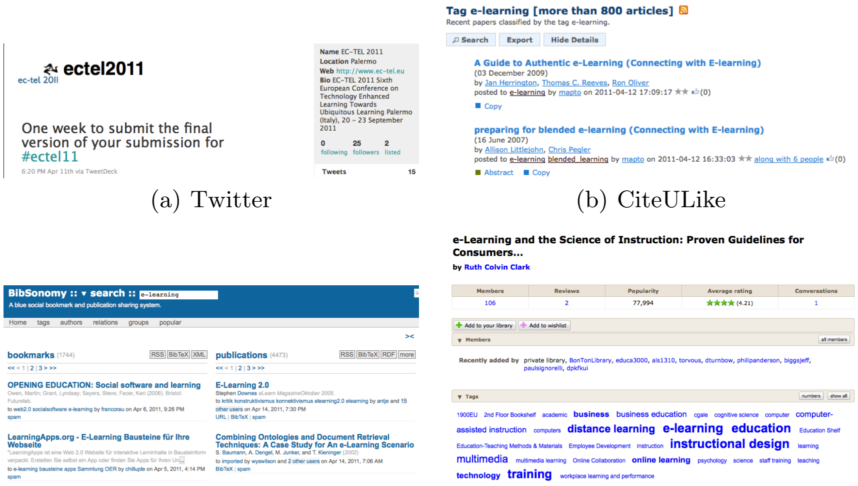
Of course, not all tweets provide us with useful information. On the contrary, the majority of tweets may be focused on day to day activities, news, sports, holidays, etc. These observations lead to two research questions:

1. Are there enough utilizable tweets to build a knowledge profile?
2. And if this is the case: How can we filter out these non-informative tweets that add noise to the profile?

Ideally, at the end of the filtering process, we would only be left with tweets that are relevant to the learner's knowledge profile.

[4] In September 2010 more than 145 million accounts were registered with the service: `http://blog.twitter.com/2010/09/evolving-ecosystem.html`
(URL last accessed in June 2011)

In order to empirically investigate these questions, we have collected tweets from people that use Twitter as well as one of three social bookmarking services, namely CiteULike[5], Bibsonomy[6] and LibraryThing[7] (examples of each service are shown in Figure 1). CiteULike and Bibsonomy are academic bookmarking services which let users bookmark scientific papers. LibraryThing is a general book management service, from which we extract the bookmarked scholarly books. For each user in our data sets, we derive the knowledge profile from one of these bookmarking services (the ground truth profile) and investigate how well it can be approximated by the user's tweets.
It should be emphasized that we solely focus on the question to what extent a knowledge profile can be constructed from Twitter data alone. We do not use the derived profiles in an application.
The rest of the paper is organized as follows. In Sec. 2 related work is presented. Then, in Sec. 3 we outline our methodology. The experiments and results are presented in Sec. 4, followed by conclusions in Sec. 5.



ectel2011
ec-tel 2011

One week to submit the final version of your submission for #ectel11
6:20 PM Apr 11th via TweetDeck

**Name** EC–TEL 2011
**Location** Palermo
**Web** http://www.ec-tel.eu
**Bio** EC–TEL 2011 Sixth European Conference on Technology Enhanced Learning Towards Ubiquitous Learning Palermo (Italy), 20 – 23 September 2011

| 0 | 25 | 2 |
|---|---|---|
| following | followers | listed |

**Tweets** 15

(a) Twitter

**Tag e-learning [more than 800 articles]**
Recent papers classified by the tag e-learning.

A Guide to Authentic e-Learning (Connecting with E-learning)
(03 December 2009)
by Jan Herrington, Thomas C. Reeves, Ron Oliver
posted to e-learning by mapto on 2011-04-12 17:09:17 ★★ (0)
Copy

preparing for blended e-learning (Connecting with E-learning)
(16 June 2007)
by Allison Littlejohn, Chris Pegler
posted to e-learning blended_learning by mapto on 2011-04-12 16:33:03 ★★ along with 6 people (0)
Abstract Copy

(b) CiteULike

**BibSonomy :: ▾ search ::** e-learning
A blue social bookmark and publication sharing system.

**bookmarks** (1744) RSS BibTeX XML
<< < 1 | 2 | 3 > >>

OPENING EDUCATION: Social software and learning
Owen, Martin; Grant, Lyndsay; Sayers, Steve; Facer, Keri (2006). Bristol: Futurelab.
to web2.0 socialsoftware e-learning by francorau on Apr 6, 2011, 9:26 PM
spam

LearningApps.org - E-Learning Bausteine für Ihre Webseite
"LearningApps ist eine Web 2.0 Website für interaktive Lerninhalte in Bausteinform verpackt. Erstellen Sie selbst ein App oder finden Sie Apps für Ihren Un...
to e-learning bausteine apps Sammlung OER by chilluple on Apr 5, 2011, 4:14 PM
spam

**publications** (4473) RSS BibTeX RDF more
<< < 1 | 2 | 3 > >>

E-Learning 2.0
Stephen Downes *eLearn MagazineOktober 2005.*
to kritik konstruktivismus konnektivismus elearning2.0 elearning by antje and 15 other users on Apr 14, 2011, 7:30 PM
URL | BibTeX | spam

Combining Ontologies and Document Retrieval Techniques: A Case Study for An e-Learning Scenario
S. Baumann, A. Dengel, M. Junker, and T. Kieninger (2002)
to imported by wyswilson and 2 other users on Apr 14, 2011, 7:06 AM
BibTeX | spam

(c) Bibsonomy

**e-Learning and the Science of Instruction: Proven Guidelines for Consumers...**
**by Ruth Colvin Clark**

| Members | Reviews | Popularity | Average rating | Conversations |
|---|---|---|---|---|
| 106 | 2 | 77,994 | ★★★★ (4.21) | 1 |

**Recently added by** private library, BonTonLibrary, educa3000, als1310, torvous, dturnbow, philipanderson, biggsjeff, paulsignorelli, dpkfkui

1900EU 2nd Floor Bookshelf academic business business education cgale cognitive science computer computer-assisted instruction computers distance learning e-learning education Education Shelf Education-Teaching Methods & Materials Employee Development instruction instructional design learning multimedia multimedia learning Online Collaboration online learning psychology science staff training teaching technology training workplace learning and performance

(d) LibraryThing

**Fig. 1.** Examples of the social Web services used in this study.

## 2 Related Work

In this section, we report on related work that discusses different aspects of Twitter: (i) the motivation for people to use Twitter and what they use it for, (ii)

[5] http://www.citeulike.org/
[6] http://www.bibsonomy.org/
[7] http://www.librarything.com/

how scholars utilize Twitter, and, (iii) how Twitter is used in learning. While a considerable number of works investigate Twitter and news (e.g., news recommendation [15], real-time event detection [19], information spread [8]), fewer works focus on Twitter as a learning aid or a source of information about a user's knowledge.

Two important questions that have been investigated by a number of researchers are why do people use Twitter and what do they tweet about. The authors in [5] developed four broad categories of tweets: daily chatter (most common use of Twitter), conversations, shared information/URLs and reported news. Naaman et al. [14] identified nine different categories: information sharing, self promotion, opinions, statements and random thoughts, questions to followers, presence maintenance, anecdotes about me and me now. Moreover, they also found that the vast majority of users (80%) focus on themselves ("Meformers"), while only a minority of users are driven largely by sharing information (”Informers”). Westman et al. [20] performed a genre analysis on tweets and identified five common genres: personal updates, direct dialogue (addressed to certain users), real-time sharing (news), business broadcasting and information seeking (questions for mainly personal information). Zhao et al. [21] interviewed people about their motivations for using Twitter; several major reasons surfaced: keeping in touch with friends and colleagues, pointing others to interesting items, collecting useful information for one's work and spare time and asking for help and opinions. These studies show that a lot of tweets are concerned with the user herself; we hypothesize that among these user centered tweets, there is also useful ones for the derivation of the learner's knowledge profile.

In [16] the Twitter posting behaviour of academics was investigated. The authors conducted a study with twenty-eight faculty, postdocs and doctoral students to determine the extent of scholars tweeting citations. About 6% of the tweets with hyperlinks were found to be citations to peer-reviewed resources. Another finding was that a large percentage (40%) of tweeted citations appear within a week of the cited resource's publication date. Tweets related to a particular scientific activity, namely conference tweets, were investigated in [9]. Here, tweets related to three conferences (identified by the conferences' official hashtags) were analysed according to how scientific information is spread on Twitter. It was found that the users mainly tag for the benefit of their own network, they do not target the wider audience. The tweets are focused on announcing future events, links to slides, publications and other related information. Although not useful to the general public, these kind of tweets are elements we consider useful in the generation of a knowledge profile from Twitter posts.

A number of Twitter studies also attempt to predict user characteristics from tweets. While we are aiming to extract knowledge profiles, Michelson et al. [12] derive topic profiles from Twitter users. In their approach, the named entities (e.g., *Barack Obama*, *David Beckham*) are extracted from tweets, they are disambiguated and linked to their corresponding Wikipedia page and then a topic profile is build. We do not follow this approach as we compare the Twitter-based profile to a ground truth profile which is based on free text (abstracts of scientific papers). In [4,13,17] elementary user characteristics are inferred from Twitter, including gender, age, political orientation, regional origin and ethnicity.

In the context of learning, first studies have begun to appear that investigate the usability of Twitter as a learning help. For instance, Borau et al. [2] utilized Twitter as part of a course on English as a foreign language for Chinese university students. The students were instructed to tweet regularly as part of actively using English. The majority of students indicated after the experiment that using Twitter made them less shy when communicating in English. McWilliams et al. [11] used Twitter as a collaborative writing tool. Twitter was also found to enhance the instructor's credibility among students, when Twitter was not only used as an academic tool, but also as a social tool by the instructor [6].

## 3 Method

Twitter users tweet short messages with up to 140 characters about anything they choose. They can be followed by other users and themselves follow users in order to receive their tweets. Tweets can be directed (*@user*) and tweets can contain hashtags (*#ectel11*).

In order to investigate how well we can derive a knowledge profile from a user's Twitter data, for each user in the data set, we perform the following procedure:

1. Calculate a knowledge profile from a scholarly bookmarking service; this is the *ground truth profile*.
   (a) Index the user's bookmarks.
   (b) Derive a term vector **b** as knowledge profile.
2. Calculate a knowledge profile from Twitter data.
   (a) Select a number of the user's tweets for indexing.
   (b) Index the selected tweets.
   (c) Derive a term vector **t** as Twitter-based knowledge profile.
3. Calculate the cosine similarity between **b** and **t**: $sim = \frac{\sum_{i=1}^{n} b_i \times t_i}{\sqrt{\sum_{i=1}^{n} b_i^2} \times \sqrt{\sum_{i=1}^{n} t_i^2}}$ where $sim \in [0,1]$ and $n$ is the number of terms in the term vector.

In the best case, that is, when the two vectors have the same direction, $sim = 1$, while at worst, $sim = 0$. The higher the similarity between the Twitter based knowledge profile and the ground truth, the better the Twitter-based knowledge profile approximates the ground truth profile.

Step 2(a), the selection of tweets for indexing, is the crucial step in our experiments - not all tweets are equally useful for the derivation of a learner's knowledge profile. Here, we investigate if it is possible to rely on simple rules to select those tweets that are useful. In the following paragraphs, we present a number of filtering options and the underlying hypotheses we have.

**Language Identification:** We cannot expect all tweets to be in English, often, tweets in English are mixed with tweets in other languages, in particular if the user is not a native English speaker. At the same time, we expect the ground truth profiles to consist largely of terms in the English language, as English is the

main language in science. Thus, excluding non-English tweets is hypothesized to reduce the noise in our data sets. A simple, yet effective, approach to language identification has been proposed in [3]. Given sets of training texts in different languages, N-grams [10] are derived for each training language and an unknown text is then assigned to the language its N-gram distribution it matches best.
**Weekdays versus Weekend:** We hypothesize that tweets made during week days are more likely to be work and study related than tweets made during the weekend. Thus, we filter out tweets that are posted on Saturday or Sunday.
**Style:** We can also consider a multitude of tweet features for training a Naive Bayes classifier [18] that determines for each tweet if it should be classified as an informative or noisy tweet with respect to the knowledge profile. A total of 19 features are derived, including whether or not the tweet is a retweet or a directed tweet, the number of words and characters in the tweet, the number of exclamation marks and question marks, the number of hashtags, the number of smileys in the tweet as well as the number of letters repeated four or more times (e.g., "oooooh" or "sooooo"). First, for a number of tweets, it is determined whether they are suitable for the knowledge profile or not (the training data). Then, the features are extracted from these tweets, a model is learnt based on the training data and this model is applied to predict whether a tweet in the test data should be included in the Twitter-based profile.
**External Documents:** Finally, instead of removing tweets, we can also include extra information. Tweets are short and although they can be informative by themselves, often a tweet contains a link and a very short expression of interest or an explanation. Thus, we also consider the external documents that are linked to as an additional potential source for the knowledge profile.

## 4 Experiments

### 4.1 Data Set Overview

As we investigate how accurately we can derive a knowledge profile from users' Twitter messages, we require a ground truth. The goal of our work is to extract user profiles that are utilizable in E-learning applications, thus we focus on the users' knowledge profiles in scholarly subjects. We decided to collect these ground truth knowledge profiles automatically, as it allows us to conduct our experiments on a larger scale. To this end, we rely on existing social Web services that are aimed at or include the organization and sharing of scholarly works. In particular, we relied upon:

- **CiteULike**: a service for organizing and sharing scientific publications.
- **Bibsonomy**: an alternative to CiteULike, that allows the bookmarking of scholarly papers as well as Web pages, and,
- **LibraryThing**: a service to catalogue books (a virtual bookshelf).

We collect the data of Twitter users, for whom we are able to link their Twitter account to one of the three bookmarking/cataloguing services listed above. We found the users in our data sets by crawling Google Profiles[8] and claimID[9], where users can list and manage their online identities on various social Web services.

Table 1 gives an overview of the collected data. The users found for each bookmarking services are treated as a separate data set. We crawled up to the latest 3150 tweets per user (Twitter limits the maximum amount of accessible tweets). In total, we found 73 users of Twitter and CiteULike, 47 users of Twitter and Bibsonomy and 122 users of Twitter and LibraryThing. The LibraryThing service is not focused on scholarly works, any book can be added. Since we are interested in the scholarly books, we only consider a subset of all bookmarks, namely those that can be found at Amazon[10] under the following categories: *Textbook*, *Science*, *Religious Studies*, *Social Sciences*, *Computer Science* and *Engineering*. The number of users in our data sets are very small compared to the total numbers of users of Twitter and the bookmarking services. Listing one's various social Web accounts on Google Profiles and claimID is voluntary and many users may either not know these services or may not feel the need nor want to publicly list their handles. The advantage of relying on Google Profiles and claimID is that the data is provided by the users themselves, thus we do not need to infer a linkage between accounts of different social Web services.

While in CiteULike and LibraryThing users bookmark publications and books only, in Bibsonomy users can bookmark publications as well as web pages.

From the CiteULike and LibraryThing services, we indexed the titles, the abstracts and descriptions (if available) as well as the tags assigned to the bookmark by the users in our data sets. In the case of Bibsonomy, we also indexed the bookmarked web pages.

Before indexing the Twitter posts, in each tweet, if applicable, the user names (*@user*) and hyperlinks (`http://bit.ly/kxreiG`) were removed. Hashtags (#educationOnline) were split according to a simple capital letter rule. Thus, a post such as "*@Tom E-learning courses start in April #educationOnline #course*" will be transformed into "*E-learning courses start in April education online course*".

All bookmarks and tweets were indexed with the Lemur Toolkit[11] with Krovetz stemming applied [7]; stopwords were removed. A user's knowledge profile is simply a vector of terms with weights according to the frequency each term occurs in the bookmark index or Twitter index. In order to avoid overestimating the similarity between the ground truth profile and the Twitter-based profile, terms that occur in more than 1% of a newspaper corpus (from the years 1995-1997) were removed; examples of removed terms are *just*, *love* and *weather*. We refrained from calculating vector elements according to TF.IDF [1], as such weights are not useful here: if a tenth of the CiteULike articles in our index for

[8] `http://profiles.google.com/`

[9] `http://claimid.com/`

[10] `http://www.amazon.com/`

[11] `http://www.lemurproject.org/`

**Table 1.** Overview of the derived data sets ($\sigma$ is the standard deviation). CiteULike and LibraryThing do not allow the bookmarking of web pages. The rows marked "Ext. URLs/Tweet" indicate the average, median, etc., number of hyperlinks in a tweet.

| | **CiteULike** | **Bibsonomy** | **LibraryThing** |
|---|---:|---:|---:|
| | **+Twitter** | **+Twitter** | **+Twitter** |
| **#Users** | 73 | 47 | 122 |
| **Average #Publications** | 490.1 | 291.3 | 63.5 |
| | ($\sigma = 689.4$) | ($\sigma = 374.3$) | ($\sigma = 82.3$) |
| **Median #Publications** | 32 | 244.0 | 162.5 |
| **Minimum #Publications** | 2.0 | 0.0 | 1.0 |
| **Maximum #Publications** | 4397.0 | 1651.0 | 460.0 |
| **Average #Webpages** | | 542.5 | |
| | | ($\sigma = 1006.1$) | |
| **Median #Webpages** | | 198.0 | |
| **Minimum #Webpages** | | 0.0 | |
| **Maximum #Webpages** | | 4038.0 | |
| **Average #Tags/Bookmark** | 3.7 | 3.2 | 1.6 |
| | ($\sigma = 2.4$) | ($\sigma = 1.6$) | ($\sigma = 1.9$) |
| **Median #Tags/Bookmark** | 3.4 | 2.9 | 0.9 |
| **Minimum #Tags/Bookmark** | 0.0 | 0.5 | 0.0 |
| **Maximum #Tags/Bookmark** | 16.5 | 7.3 | 7.4 |
| **Average #Twitter Posts** | 1607.0 | 808.7 | 1909.4 |
| | ($\sigma = 1235.1$) | ($\sigma = 1011.2$) | ($\sigma = 1182.8$) |
| **Median #Twitter Posts** | 3150.0 | 3095.0 | 3150.0 |
| **Minimum #Twitter Posts** | 23.0 | 1.0 | 1.0 |
| **Maximum #Twitter Posts** | 3150.0 | 3150.0 | 3150.0 |
| **Average #Ext. URLs/Tweet** | 0.5 | 0.5 | 0.4 |
| | ($\sigma = 0.3$) | ($\sigma = 0.3$) | ($\sigma = 0.3$) |
| **Median #Ext. URLs/Tweet** | 0.4 | 0.5 | 0.3 |
| **Minimum #Ext. URLs/Tweet** | 0.0 | 0.0 | 0.0 |
| **Maximum #Ext. URLs/Tweet** | 1.8 | 1.0 | 1.5 |

example would include the term *genetics*, it would receive a low weight, although it may actually represent the user's knowledge profile very well.

Examples of the top weighted terms of typical user profiles derived from the four social Web services are shown in Table 2. Note, that they are all from different users; shown are the stemmed terms. The CiteULike profile clearly focuses on bioinformatics and genetics, whereas the Bibsonomy profile contains terms typical for the Semantic Web. This is a general trend we found across these two services in our data sets: CiteULike is frequented by users coming from the biomedical domain, while Bibsonomy is used by users whose profile indicates work in Computing Science and related areas. The LibraryThing profile shown is mixed: on the one hand, it contains terms that indicate knowledge in areas of computer science (*visualize*, *internet*, *web*), but on the other hand it also contains terms such as*obama* and*barack*, indicating an interest in the political domain. Since the LibraryThing service lets users add books they have read (or might read), instead of scientific papers, it is possible to a lesser degree to make a distinction between a user's general interest and his scholarly knowledge. Furthermore, the categories we rely on to filter academic books play a role here as well, as books may belong to different categories. The Twitter profile in the last column shows that this user has tweeted the most about current news.

**Table 2.** The top weighted terms of example user profiles drawn from each data set.

| **CiteULike** | **Bibsonomy** | **LibraryThing** | **Twitter** |
|---|---|---|---|
| connotea | ontology | web | rt |
| csb | stlab | navigation | wikileak |
| interaction | web | findable | assange |
| bacillu | semantic | obama | cablegate |
| gene | workshop | morville | libya |
| cell | owl | barack | leak |
| stochastic | dns | interface | guardian |
| yeast | descriptionsandsituations | visualize | google |
| genetic | dolce | internet | twitter |
| biology | ontologydesign | designer | wiki |

### 4.2 Results

We first present the results of the baseline (how well can the knowledge profile be approximated by utilizing all tweets of a user?) and the results of the upper bound (if we have an oracle, that tells us which tweets are the right ones, what is the best possible knowledge profile we can achieve?). Then, the results of the tweet selection and expansion experiments are reported. We will show, that Twitter is a useful source for deriving knowledge profiles, though predicting which tweets aid in profiling is a difficult task.

**Baseline & Upper Bound** The baseline is derived by calculating the similarity between the Twitter-based profiles and the ground truth (bookmarking service based) profiles. We report the mean, median, minimum and maximum cosine similarity across the users of each data set. The results are reported in Table 3. CiteULike and Bibsonomy reach an average similarity of 0.18 and 0.2, respectively, while for the LibraryThing data set the similarity is considerably lower (0.07). This shows, as expected, that profiles based on all tweets, on average, are not suitable for deriving a learner's knowledge profile. However, when we consider the maximum cosine similarity, that is, the similarity reached by the user whose tweets match the ground truth profile the most, in the CiteULike data set, the similarity is greater than 0.9, whereas in the other two data sets the similarity reaches $\approx 0.5$. Thus, indeed there are some users, whose tweets are very much related to their professional life, while for most users, those tweets are either not existing or hidden among the non-informative tweets.
In order to investigate if users simply do not tweet about their study or work life or if it is indeed a case of overbearing non-informative tweets, we experimentally derived the highest cosine similarity possible with our model and our data. We conducted the following experiment: at step $k = 0$, we start with a set $T$ which contains all tweets $t_1, ..., t_m$ of a user and an empty set $S$. We then iteratively add tweets to $S$: in each step $k$, we add the tweet $t_x$ to $S$ such that $S \cup t_x$ together form the Twitter-based profile having the highest cosine similarity with the ground

truth profile. Specifically, in the first step ($k = 1$), the tweet $t_i$ (among all tweets in $T$) is selected which itself has the highest cosine similarity with the ground truth profile. Tweet $t_i$ is added to the empty set $S$ and removed from set $T$. In step $k = 2$, the next selected tweet from $T$ is the one that together with tweet $t_i$ from set $S$ forms the profile that is most similar to the ground truth profile. And so on for $k = 3, 4, ..., m$. At step $m$, the set $T$ is empty and set $S$ contains all tweets. At each step $k$, we record the cosine similarity the tweets in $S$ reach with respect to the ground truth profile. The greedy upper-bound is then the maximum similarity we record across all $k$[12]. The upper-bound results for each data set, are also reported in Table 3. On average, if the "right" tweets are selected to represent the user's knowledge profile, the average cosine similarity reaches between 0.4 (LibraryThing) and 0.6 (CiteULike), which are substantial improvements over the baselines. The minimum similarity is still low; there are Twitter users that offer very few or no suitable information in their tweets. On the other hand, the maximum similarity with the ground truth reaches 0.8 or higher across all data sets. These results show, that if we would be able to select the right tweets, we could derive useful knowledge profiles from Twitter.

**Table 3.** Baseline results and greedy upper bound. Reported is the mean (standard deviation), median, maximum and minimum cosine similarity between the Twitter-based profile vectors and the bookmarking service-based profile vectors across the users of each data set.

| **Data Set** | | **Mean** | **Median** | **Min.** | **Max.** |
|---|---|---|---|---|---|
| **CiteULike** | baseline | 0.176 ($\sigma = 0.171$) | 0.143 | 0.002 | 0.933 |
| | upper-bound | 0.608 ($\sigma = 0.241$) | 0.654 | 0.008 | 0.994 |
| **Bibsonomy** | baseline | 0.203 ($\sigma = 0.130$) | 0.192 | 0.003 | 0.550 |
| | upper-bound | 0.545 ($\sigma = 0.253$) | 0.555 | 0.006 | 0.919 |
| **LibraryThing** | baseline | 0.075 ($\sigma = 0.070$) | 0.059 | 0.000 | 0.473 |
| | upper-bound | 0.412 ($\sigma = 0.185$) | 0.411 | 0.022 | 0.838 |

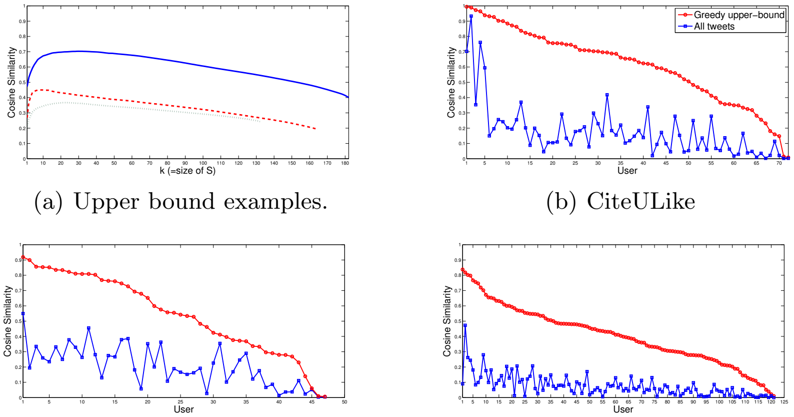
In Figure 2(a) we plot the development of the cosine similarity of set $S$ for the tweets of three random users. Although the absolute cosine similarity differs, the process is similar across all of them: initially, adding tweets to set $S$ increases the cosine similarity, but early on a peak is reached (the reported upper-bound) and adding more tweets reduces the cosine similarity again. Across the data sets, the average, minimum and maximum number of tweets at the peak are reported in Table 4. Thus, although for the majority of users we have 3150 tweets in our data set, the highest similarity with the ground truth is reached after $30 - 40$ tweets.

[12] The upper bound is "greedy" due to the iterative process. It is an approximation of the true upper-bound, which would require calculating all possible combinations of tweets in $S$, which is computationally not feasible.

**Table 4.** Mean, minimum and maximum number of tweets at the greedy upper-bound.

| | **Mean** | **Minimum** | **Maximum** |
|---|---|---|---|
| **CiteULike** | 42.1 | 1 | 229 |
| **Bibsonomy** | 31.3 | 1 | 113 |
| **LibraryThing** | 35.8 | 1 | 133 |

To provide a better impression of the difference between the baseline and upper-bound for each individual user, consider the plots in Figure 2(b), 2(c) and 2(d). Here, in each plot, we sorted the users in the data set according to the cosine similarity of their upper-bound from high to low and plotted the corresponding baseline. It is evident, that across all data sets the gained improvements are large. It is also apparent, that even if the baseline similarity is low, if the right tweets are selected, a very good knowledge profile can be generated.



(a) Upper bound examples.

(b) CiteULike

(c) Bibsonomy

(d) LibraryThing

**Fig. 2.** Examples of the upper bound over a range of $k$ are shown in Fig. 2(a). The cosine similarity for the baseline and upper bound across all users of each data set are plotted in Fig. 2(b)-2(d).

**Tweet Filtering & Expansion** In the next step, we implemented the heuristics for tweet filtering and expansion introduced in Sec. 3. First, we filter out all tweets and bookmarks that are not in English. While the number of Twitter posts decreased by 14% when filtering out non-English posts, the three bookmarking services had fewer entries identified as non-English (CiteULike 1%, Bibsonomy 8%, LibraryThing 0.5%). In a second experiment, we removed all tweets posted at the weekend from the Twitter profiles. In a third experiment, we expanded the Twitter index by including documents that are linked from tweets. Finally, we built a Naive Bayes classifier and performed 5-fold cross validation: each data set was split into five equal parts and four parts were used for training the classifier and one was used for testing. This procedure was repeated 5 times (each time, a different part was held out for testing) and the results on the test data were averaged. The labels for the training data were derived automatically: tweets that

have a high cosine similarity with the ground truth profile were labelled as informative (to be selected) while tweets with a similarity of ≈ 0 with the ground truth profile were labelled non-informative (not to be selected). The classifier was applied on the test data and each tweet was classified as informative or not and only the informative tweets were included in the Twitter-based knowledge profile.
The results of all experiments are shown in Table 5. Underlined entries indicate an improvement over the respective baseline (Table 3). Selecting only tweets posted during weekdays does not yield improvements over the baseline for any data set when considering the mean, our most important measure. Excluding non-English tweets has a positive effect on the CiteULike and LibraryThing data sets, though only marginally. Expanding the Twitter based profile by including documents whose links were tweeted has a drastic effect on the Bibsonomy data set: while the baseline mean cosine similarity is 0.20, the mean cosine similarity of the expanded profile is 0.35, a 75% increase. Notable is also the increase in the maximum of the CiteULike data set for this experiment: while in the baseline, the maximum similarity is 0.55, in the expanded profile, it reaches 0.85, a 55% increase. We strongly suspect that this result is due to the user group that we found to mostly make up our Bibsonomy user set: users bookmarking papers in areas of computer science. Due to the nature of the field, there is a lot of relevant information on the Web and thus hyperlinks posted on Twitter by such users may often refer to aspects of computer science.

**Table 5.** Overview of the cosine similarity when performing tweet selection and including linked external documents.

| **Filtering** | **Data Set** | **Mean** | **Median** | **Minimum** | **Maximum** |
|---|---|---:|---:|---:|---:|
| **English Only** | CiteULike | $\underline{0.184}$ ($\sigma = 0.178$) | $\underline{0.145}$ | 0.002 | $\underline{0.942}$ |
| | Bibsonomy | 0.202 ($\sigma = 0.129$) | 0.188 | $\underline{0.004}$ | 0.536 |
| | LibraryThing | $\underline{0.077}$ ($\sigma = 0.070$) | $\underline{0.060}$ | 0.000 | $\underline{0.482}$ |
| **Weekdays Only** | CiteULike | 0.173 ($\sigma = 0.167$) | 0.143 | 0.000 | 0.921 |
| | Bibsonomy | 0.200 ($\sigma = 0.134$) | 0.190 | $\underline{0.004}$ | 0.549 |
| | LibraryThing | 0.074($\sigma = 0.070$) | 0.058 | 0.000 | $\underline{0.477}$ |
| **Including External Documents** | CiteULike | 0.169 ($\sigma = 0.118$) | $\underline{0.167}$ | $\underline{0.005}$ | 0.474 |
| | Bibsonomy | $\underline{0.350}$ ($\sigma = 0.254$) | $\underline{0.332}$ | $\underline{0.011}$ | $\underline{0.846}$ |
| | LibraryThing | $\underline{0.079}$ ($\sigma = 0.068$) | $\underline{0.064}$ | 0.000 | 0.339 |
| **Naive Bayes** | CiteULike | $\underline{0.192}$ ($\sigma = 0.168$) | $\underline{0.151}$ | $\underline{0.003}$ | $\underline{0.936}$ |
| | Bibsonomy | 0.195 ($\sigma = 0.128$) | 0.170 | 0.000 | $\underline{0.557}$ |
| | LibraryThing | $\underline{0.081}$ ($\sigma = 0.070$) | $\underline{0.064}$ | 0.000 | $\underline{0.489}$ |

The results of the Naive Bayes classifier are inconsistent, we found the largest improvement, 9%, over the baseline in the CiteULike data set (mean cosine similarity); however, in the Bibsonomy data set, the baseline outperformed the classifier based tweet selection. This implies, that while predicting which tweets are concerned with aspects of a user's work or study is to some degree influenced

by the style of a tweet, it should not be the only source of information. The tweet content and the Twitter network structure of the user are likely to play a significant role as well.
Across all experiments, the LibraryThing data set performs less well than the CiteULike and Bibsonomy data sets. It will require further investigation to determine the differences and similarities between them. One potential explanation may be, that the description of the bookmarked books are often short and may not always contain the key concepts. The bookmarked scientific publications in CiteULike and Bibsonomy on the other hand do mostly also contain the abstract of the work, which by its nature contains the keywords and topic words that we expect in the profile vectors.

## 5 Summary and Future Work

In this work, we set up a framework that acts as a testbed for further exploration of learner profile gathering on the social Web. We investigated how well a profile built from a user's tweets can approximate a user's knowledge profile. We offered a methodology of how to collect relevant data sets automatically, by considering users that explicitly link their Twitter account and a scholarly bookmarking service account together. We found that indeed a large number of users tweet not only about news, sports, etc. but also about aspects of their professional life. By determining the greedy upper-bound between the ground truth profiles and Twitter-based profiles, we could show that if the right tweets were selected, a good approximation of the ground truth profile is possible for the majority of users. We further found that for different user groups, different aspects of Twitter posts are useful, in particular the Bibsonomy data set, which includes many users with knowledge related to computer science, profited considerably from the inclusion of documents that were linked to in Twitter posts. Finally, we observed that predicting which tweets to include in the Twitter based profile is a difficult task.
These results leave a lot of potential for future work. A limitation of our work, that we have not yet adressed is the question of whether the users in our data set resemble the average Twitter users. Based on the genres, types of tweets and their distribution [14, 20], we will conduct a qualitative analysis of a set of random tweets in our data sets. Additionally, we will further investigate the tweet selection problem by increasing the number of features, but also by taking Twitter's network structure into account. Finally, we also plan to increase the data set size further, and to investigate different user groups within each data set such as professionals versus students, males versus females and user groups of different geographic origin.

## References

1. R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., 1999.

2. K. Borau, C. Ullrich, J. Feng, and R. Shen. Microblogging for language learning: Using twitter to train communicative and cultural competence. *Advances in Web Based Learning – ICWL 2009*, pages 78–87, 2009.
3. W. Cavnar and J. Trenkle. N-gram-based text categorization. In *SDAIR '94*, pages 161–175, 1994.
4. B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *CHI '11*, pages 237–246, 2011.
5. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.
6. K. Johnson. The effect of Twitter posts on students perceptions of instructor credibility. *Learning, Media and Technology*, 36(1):21–38, 2011.
7. R. Krovetz. Viewing morphology as an inference process. In *SIGIR '93*, pages 191–202, 1993.
8. K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *ICWSM '10*, pages 90–97, 2010.
9. J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how Twitter is used to widely spread Scientific Messages. In *WebSci10: Extending the Frontiers of Society On-Line*, 2010.
10. C. D. Manning and H. Schütze. *Foundations of statistical natural language processing.* MIT Press, Cambridge, MA, USA, 1999.
11. J. McWilliams, D. Hickey, M. Hines, J. Conner, and S. Bishop. Using Collaborative Writing Tools for Literary Analysis: Twitter, Fan Fiction and The Crucible in the Secondary English Classroom. *Journal of Media Literacy Education*, 2(3):238–245, 2011.
12. M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *AND '10*, pages 73–80, 2010.
13. A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the Demographics of Twitter Users. In *ICWSM '11*, 2011.
14. M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: message content in social awareness streams. In *CSCW '10*, pages 189–192, 2010.
15. O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *RecSys '09*, pages 385–388, 2009.
16. J. Priem and K. L. Costello. How and why scholars cite on twitter. In *ASIS&T '10*, pages 75:1–75:4, 2010.
17. D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *SMUC '10*, pages 37–44, 2010.
18. I. Rish. An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, pages 41–46, 2001.
19. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW '10*, pages 851–860, 2010.
20. S. Westman and L. Freund. Information interaction in 140 characters or less: genres on twitter. In *IIiX '10*, pages 323–328, 2010.
21. D. Zhao and M. B. Rosson. How and why people twitter: the role that microblogging plays in informal communication at work. In *GROUP '09*, pages 243–252, 2009.