Supporting collaborative learning with tag recommendations: a real-world study in an inquiry-based classroom project

Simone Kopeinik
ISDS
Graz University of Technology
Graz, Austria
simone.kopeinik@tugraz.at

Elisabeth Lex
ISDS
Graz University of Technology
Graz, Austria
elisabeth.lex@tugraz.at

Paul Seitlinger
School of Educational Sciences
Tallinn University
Tallinn, Estonia
paul.seitlinger@tlu.ee

Dietrich Albert
ISDS
Graz University of Technology
Graz, Austria
dietrich.albert@tugraz.at

Tobias Ley
School of Educational Sciences
Tallinn University
Tallinn, Estonia
tley@tlu.ee

ABSTRACT

In online social learning environments, tagging has demonstrated its potential to facilitate search, to improve recommendations and to foster reflection and learning.Studies have shown that shared understanding needs to be established in the group as a prerequisite for learning. We hypothesise that this can be fostered through tag recommendation strategies that contribute to semantic stabilization. In this study, we investigate the application of two tag recommenders that are inspired by models of human memory: (i) the base-level learning equation BLL and (ii) Minerva. BLL models the frequency and recency of tag use while Minerva is based on frequency of tag use and semantic context. We test the impact of both tag recommenders on semantic stabilization in an online study with 56 students completing a group-based inquiry learning project in school. We find that displaying tags from other group members contributes significantly to semantic stabilization in the group, as compared to a strategy where tags from the students' individual vocabularies are used. Testing for the accuracy of the different recommenders revealed that algorithms using frequency counts such as BLL performed better when individual tags were recommended. When group tags were recommended, the Minerva algorithm performed better. We conclude that tag recommenders, exposing learners to each other's tag choices by simulating search processes on learners' semantic memory structures, show potential to support semantic stabilization and thus, inquiry-based learning in groups.

LAK '17 March 13–17, 2017,Vancouver, BC, Canada

ACM ISBN 978-1-4503-4870-6/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3027385.3027421

CCS Concepts

•Information systems → Social tagging systems; Recommender systems; •Applied computing → Collaborative learning; •Computing methodologies → Cognitive science;

Keywords

Semantic stabilization; Personalized tag recommendations; Cognitive user models; Base-level learning equation; Minerva; Real-world testing; Inquiry-based learning; Technology enhanced learning

In open and social learning environments, students need to construct knowledge in a self-directed manner in a social context. In inquiry-based learning (IBL), for example, students are encouraged to collect and retrieve information and create new content, which is continuously uploaded to the learning environment. In these settings, tagging has demonstrated its potential to enrich awareness and reflection of students. An empirical study conducted by Kuhn et al. [26] indicates that tagging supports learning in IBL by helping students organize, retrieve and reflect upon the content of learning resources they found on the Web (e.g., learning videos) or generated themselves (e.g., blog entries). Mediated by tagging, these activities become inherently social activities (e.g., [14]) as the tag vocabulary, on which a student draws to organize and reflect on resources, emerges not only from personal tag choices, but also from those of others. Students of an IBL setting can thus be expected to benefit from semantic stabilization (e.g., [40]) – a phenomenon that becomes manifest in an increasing convergence in choosing tags for particular ranges of topics: the more stable the currently evolved tag vocabulary is, the more helpful it should be to share own and exploit others' search results (i.e., Web resources). And indeed, a study of Ley and Seitlinger [27] revealed that students tend to acquire more knowledge about domain concepts, if they act in groups that exhibit relatively higher levels of semantic stabilization. We therefore conclude that IBL can be supported by processes helping students in achieving convergence in the naming of learning concepts and development of a more stable tag vocabulary. One strategy to drive such processes is to apply tag recommendation mechanisms that stabilize tagging habits by displaying tags already used in the past (e.g., [12]).

Within technology enhanced learning (TEL) research, recommendation mechanisms are part of adaptation or personalization strategies that support students in their individual learning processes [10]. Recommendation mechanisms, extract and draw on relevant data from learning traces, leveraging learning analytics [11, 7]. This is one way of tackling the often criticized lack of support for the self-organization of learning content in TEL environments [3].

A tremendous number of recommendation approaches have been suggested in recent years [20]. However, research mainly focuses on the recommendation of learning resources, peers and activities [30, 10]. TEL research on tag recommendation mechanisms and its potential for learning is still widely unattended [21].

Problem. The aim of this work is to test the impact of tagging recommendation mechanisms on semantic stabilization in an online IBL setting. With regard to results presented in Font et al. [12] it is fair to assume that an increased awareness of peer learners' tag choices will promote the development of a common terminology. In particular, we are interested to measure the performance of tag recommendations in two settings: personal (P), where students receive tag recommendations based on their personal tagging history and collective (C), where students receive tag recommendations based on the collective tagging history of their learning group.

Additionally, our work is motivated by a more technical stance: When selecting a proper tag recommendation strategy, TEL specific requirements need to be taken into account. For instance, in TEL scenarios, data is typically of a sparse nature [39]. Furthermore, sensitivity to an algorithm's complexity is crucial when calculating real time recommendations on limited computing resources [31].

Approach and Methods. A very simple, though relatively effective, tag recommendation strategy is the Most Popular (MP) algorithm [19, 22]. We however assume that a frequency-based, computationally simple recommendation strategy may be even more successful, if it is grounded on a thorough understanding of how humans process information. Our hypothesis is that in online social learning environments, semantic stabilization can be fostered by cognitively inspired tag recommendation approaches. Offline data studies have indicated that the modelling of cognitive processes underlying tagging habits leads to an increased accuracy of recommendations [38]. However, offline data studies are limited to evaluating the prediction of user behaviour. In our previous work [24, 22], we have intensively investigated the suitability of two tag recommendation approaches via offline studies [22]: the first of these, known as BLL implements the Base Level Learning Equation [1], which models the frequency and recency of past tag use. The second algorithm, known as Minerva [18, 36], incorporates tag use frequency as well as semantic context. Both approaches aim to imitate cognitive processes of retrieving words from memory.

In the present work, we study the performance of the algorithms in an online, real-world scenario to explore whether the promising results from offline data studies generalize to online environments. To this end, we carried out a field study in which students used an online IBL environment in a realistic school context for a duration of about four weeks. Our aim is to investigate the effectiveness of the two cognitively inspired recommendation mechanisms BLL and Minerva that mimic student's tagging behaviour, taking into account either temporal or semantic context.

Contributions. Our contributions are twofold: First, we investigate the question of whether semantic stabilization – a socio-cognitive process supporting individual learning in an online IBL environment [27] – can be supported by tag recommendation mechanisms that have been developed and tested previously in offline studies on a variety of data sets.

Second, we systematically vary two variables underlying the design of these recommenders in order to derive more precise and practical design implications for specific learning settings. The first variable is the vocabulary, from which the algorithm selects the tags and which can either be the learner's personal (P) or the collective vocabulary of the whole group of learners (C). The second variable is the type of information the algorithm takes into account to estimate the current probability of a tag being retrieved from the learner's memory. While MP only considers a tag's usage frequency (baseline), our two cognitively inspired algorithms extend this approach by the information of recency of usage (BLL) and the extent to which a tag matches the current (i.e., the resource's) semantic context (Minerva).

The results indicate that the application of recommenders using collective tagging traces fosters semantic stabilization in collaborative learning settings. The consideration of frequency and semantic context further contributes to the adequacy of tagging recommendations. In respect to individual learning we find that a frequency and recency based approach (BLL) performs best.

1. RELATED WORK

At present, we identify three main lines of research related to our work: tagging as a support in learning, semantic stabilization in social learning systems and tag recommendation approaches in the context of TEL.

1.1 Tagging as a Support in Learning

Due to the growing quantity of learning resources and learning data available in digital learning repositories and on the web [10], learners often struggle with the organization, the retrieval and even the awareness of relevant learning content [9, 2].

Tagging, as a simple mechanism to annotate resources individually or socially, has demonstrated its potential to facilitate search, to improve recommendations and to foster reflection and learning on the Web precisely as in technology-enhanced learning environments [41, 17]. For instance Kuhn et al. [25] investigated the effect of learning item annotation in the context of IBL and found that tagging encourages students to reflect upon retrieved learning contents. Moreover, Bateman and Brusilovsky [3] argue that according to Bloom's taxonomy of learning [6], learner's engagement in the tagging process fosters the development of a metacognitive level of knowledge, and hypothesize that the evaluation of peer learners tags might even lead to a deeper level of learning.

Also, tagging in open and social learning settings can be applied as an alternative to the unpopular, since resource intensive, description of learning items through the adding of metadata that is typically done by expert users [3]. Contrary to indexing mechanisms with controlled vocabularies, tagging allows for unrestricted extension of verbalism: social tagging systems are not bound to the use of predefined language or terminology, but its classification vocabulary grows with its users' interactions. This entails advantages, such as the support of an adaptive level of granularity, but also challenges such as the lack of a coherent and useful tag vocabulary [32]. Along these lines, research [26, 27] indicates that students seek assistance in the tagging process, regarding two aspects: (a) the take up of the process and therefore, the finding of initial vocabulary and (b) the achievement of a semantically stable vocabulary amongst their learning peers.

1.2 Semantic Stabilization in Social Learning Systems

Individual user's tagging of items shows great potential in the organization of knowledge within and across information systems [29]. However, the usefulness of such annotations is conditioned by the development of a shared terminology that leads to a meaningful description of resources [40]. The attainment of an implicit consensus on a collective vocabulary within a group, which is stable over time and in meaning, is called semantic stability [35]. In this work, we use the notion of semantic stabilization not to refer to a point in time, at which such consensus is reached and remains stable thereafter, but, we use it and a simple measure thereof (see Section 3.1) to merely characterize and compare the evolution of convergence in tag choices of two groups of students over a short period of time (few weeks).

Fu et al. [15] shows that throughout the learning process (e.g., the exploration of knowledge) semantic structures of users in a social tagging environment assimilate. Thus, learners are influenced by the tagging behaviour of their peers. Other research assumes a mutual influence between learners' internal knowledge representation and the tagging vocabulary that emerges in the social information system, in which they interact [13]. Ley et al. [27] investigates these dynamics and finds a positive influence of semantic stabilization on individual learning. Following this, we do not assume stabilization to be a prerequisite for learning but only that it provides some helpful structure for individual learning activities and is therefore conducive to individual learning gains [27].

1.3 Tag Recommendation Approaches in the Context of TEL

Despite the reported potential of tagging and students' demand for tagging support, the study of tag recommendation mechanisms has been widely unattended in TEL research [21]. Noteworthy are some initial attempts that aim to fill this gap: Diaz et al. [8] investigated the automated tagging of learning objects utilizing a computationally expensive variant of Latent Dirichlet Allocation [5] and evaluated the tagging predictions in a user study. In Niemann et al. [33], an approach to automatically tag learning objects based on their usage context was introduced. It shows promising results towards the retrospective enhancement of learning object meta-data. However, their approach cannot be used in online settings as it is based on context information of resources that is extracted from user sessions. Kopeinik et al. [22] presents an offline data study comparing a variation of tag recommendation strategies on six TEL data sets. The present work builds upon this study, since it encompasses algorithms that are applicable in runtime-sensitive online environments such as often found in educational settings. Moreover, their results have shown that simple recommendation mechanisms based on the Base Level Learning Equation (BLL) and Most Popular (MP) outperform other state-of-the-art algorithms.

2. EXPERIMENTAL SETUP

To test our hypothesis that in online social learning environments, semantic stabilization can be fostered by cognitively inspired tag recommendation strategies, we implemented a real-world evaluation in the context of high school biology lessons, engaging students in IBL projects. To this end, we used an online environment for open social inquiry-based learning. IBL itself is very well-suited to the purpose of a collaborative tagging study as throughout the learning process, students are constantly challenged to find, create, upload and share content. In the course of the study, four secondary school classes with students at the age of 15 to 17 used a dedicated social learning environment to work on their biology projects.

2.1 Procedure

Prior to the first lesson, students and their parents were presented with the goals and benefits of using IBL and the online learning environment, as well as with the aims of the study. Afterwards, parents and students were asked for their consent in written and verbal form, respectively. Throughout the study students' participation was not obligatory, in either the platform or in tagging and did not contribute to their grading.

For the purpose of the study, students of each class were divided in two groups per class, which led to groups of 9 to 18 students, depending on class size. In the first two lessons, students were introduced to the online learning environment (see Section 2.3) and the concepts of IBL. Then, each group used the virtual learning environment during at least eight school lessons over a period of four weeks or longer to complete an IBL project. The teacher provided each of the classes' learning groups with similar learning content and learning tasks and acted in a supporting role. The variation between groups is constituted by the nature of tag recommendations. Tag recommendations of one group are based on individual user's personal tag data, whereas the second group's tag recommender draw on the group's collective tagging traces. According to the group, tag recommendation strategies were randomly selected either from the personal or the collective pool of recommendation strategies, as illustrated in Table 1. Each student was provided with a tablet computer available during class. Students were encouraged to use the tagging functionality when creating or uploading new content, and were also provided with information in verbal and written form, on how to do this.

2.2 Study Design

We investigated the suitability of BLL and Minerva for facing the challenges of real-world learning settings. The resulting data sample consists of N=56 students with an age ranging from 15 to 17 years. As summarized in Table 1, the independent variables formed a 2 (Vocabulary: Personal vs. Collective; between-subjects) by 3 (Algorithm: MP vs. BLL vs. Minerva; within-subjects) design. In addition we

Table 1: Students of each class were separated in two groups and consequently received either tag recommendations based on their personal tagging history (P) or based on the collective tagging traces of their inquiry group (C). Condition C was complemented by the mixed approach BLL_U+ MP_G.

Vocabulary	Algorithm			
Personal (P)	MP_U	BLL_U	$Minerva_U$	
Collective (C)	MP_G	BLL_G	$Minerva_G$	BLL_U+ MP_G

consider BLL+MP for the collective vocabulary condition. This recommendation approach is of particular interest, as in offline studies on TEL data sets it clearly outperformed remaining algorithms [22]. The dependent variables were semantic stabilization and recommender accuracy (see Section 3).

2.3 Environment

The study was implemented in the collaborative online learning environment weSPOT[1]. weSPOT is a European research project, and stands for Working Environment with Social and Personal Open Tools for IBL. In the course of the project, a theoretical framework and corresponding TEL tools for science learning and teaching have been developed. The tool set aims to support teachers in the application of IBL as a classroom activity [37].

The weSPOT platform, or more concretely, the weSPOT inquiry space guides students through the inquiry cycle, which models the scientific inquiry process in six phases: Question/Hypothesis, Operationalisation, Data Collection, Data Analysis, Interpretation/Discussion and Communication. Each phase further includes dedicated activities as discussed in Protopsaltis et al. [34].

Figure 1 exhibits the weSPOT inquiry space that implements the six IBL phases, providing individual tabs for each phase (1). The phase-tabs further include widgets (2) that enable the students to carry out activities relevant to a specific phase. Reflection and support tools such as a learning analytics dashboard, an open user model and a recommendation interface are provided in the platform's side panel (3).

The weSPOT space supports collaborative learning in defined groups. Each student group is thus provided with a sub-environment that forms an inquiry space addressing a specified research interest. Teachers take on a supportive and administrating role. In the platform, they are provided with a configuration interface, for designing inquiry spaces by selecting phases (tabs) and activities (widgets) that suit the purpose of their student's inquiry projects. Teachers also add students and initial learning content to the group environment.

While students work on their inquiry projects, they engage in activities that typically create content by, e.g., posting questions, starting or contributing to discussions or by uploading documents and pictures. These and other learning activities are tracked, saved and fed into different user profiles, to be later used in learning analytic diagrams, to issue badges and to provide personalized recommendations of learning resources and tags.

Technical Insights. The core of the weSPOT environment is an online platform which is based on elgg[2]. Elgg is an open source social networking engine that is extendable via plugins and follows a MVC (Model-View-Controller) pattern, which makes it convenient to extend. When a user enters content (e.g., question, hypothesis, file or discussion entry) to an inquiry, this happens through an input form which includes a "tag view". The tag recommendation plugin is an extension of this "tag view" and adaptively suggests tags to users. The tag view (thus also our tag recommendation functionality) is by default included in all plug-ins that allow users to create content, for instance in discussions, file uploads or blog entries. Figure 2 shows such an input form with our recommendation plug-in embodied as marked by the orange frame. Recommendations are calculated in a backend web-service component, on the basis of the randomly selected recommendation strategy. Following common practice in social tagging systems, we set the number of tag recommendations to five. However, due to the cold start of the user and group environment, fewer tags may be presented if fewer tags are available. Learners can either select from the suggested tags by clicking on the tag or they can manually enter their own tags.

Tagging Interface. Figure 2 shows an extended version of the environment's standard input form. The tag recommendation plugin that extends the form is marked with an orange frame.

Within this study, the annotation process consisted of two steps: firstly, the selection of semantic features (attributes) from a provided dropdown menu (1): the attributes were drawn from the inquiry's domain model which has been provided by the teacher. Further information on the domain model and related tools can be found in Bedek et al. [4]. Secondly, the assignment of tags: after the student closed the dropdown menu, tag recommendations (3) appeared just

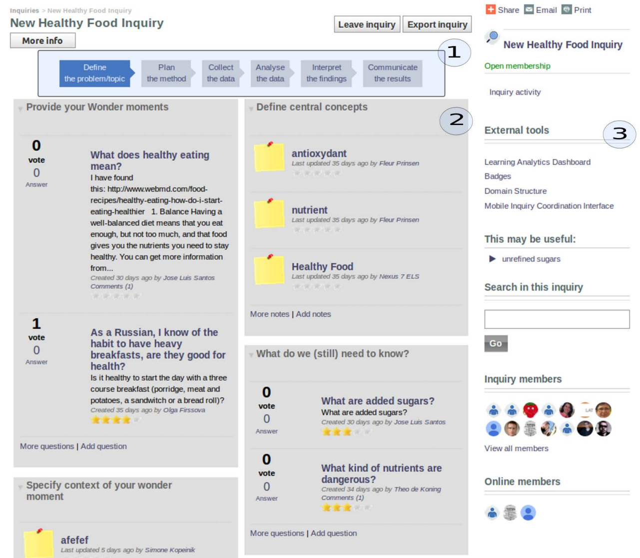


Figure 1: The collaborative online learning platform: weSPOT IBL space. (1) shows the IBL phases that are depicted as one tab each, (2) the widgets in one tab and (3) the side panel with external (supporting) tools and group information.

[1]http://inquiry.wespot.net/

[2]https://elgg.org/

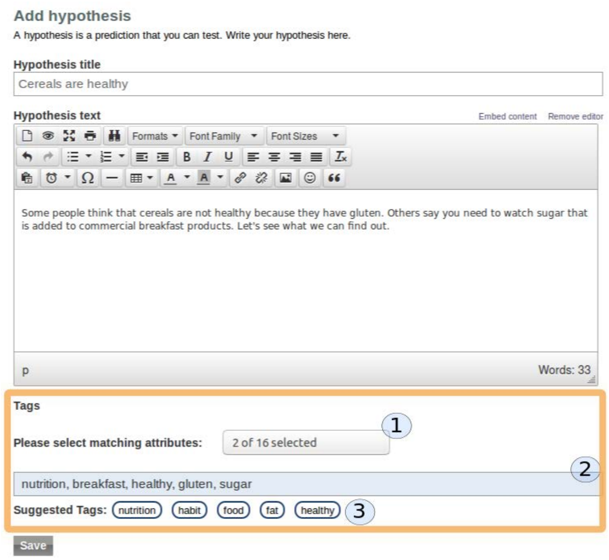


Figure 2: The standard elgg input form extended by our tag recommendation plugin (marked with the orange frame). After choosing relevant semantic features, students can either select from recommended tags (3) by clicking on the selected item or enter their own tags in the text field (2).

below the tags input text field (2). Students could either select from these recommendations or add their own tags manually.

2.4 Learning Analytics Data Collection

The data sets used in this study were collected on a dedicated log data server, from which we extracted the eight inquiry groups that participated in our experiment. All groups consisted of students attending a high school in Graz and worked on the projects in the course of biology classes, on altogether four different research topics. As one setting took place in the course of an extra-curricular specialisation, ten students participated twice in the experiment. The data was collected over a period of three school semesters (i.e., spring 2015 till summer 2016).

Although students were provided with initial instructions on the tagging interface and the tagging process itself, a relatively large number of students did not tag at all, or provided tags in unusual ways. Consequently, we manually pre-filtered the data sets by mainly excluding posts with tags in form of sentences or tags concatenated with special characters. Students with no remaining posts were also excluded from the data sets, which led to the data samples given in Table 3a.

To evaluate semantic stability, which we measured in tag growth (TG), we selected the class which created the highest number of posts in both inquiry groups. We consider this sample as the most significant to our investigation. To measure the recommendation accuracy (RA), we subsume all samples under the independent study variable *Vocabulary*. The resulting data set properties are presented in Table 3b.

3. EVALUATION MEASURES

Our study design treats semantic stability and recommendation accuracy as dependent variables. In this section, we give insight into the evaluation of both variables.

3.1 Semantic Stabilization

As summarized in Wagner et al. [40], a multitude of metrics is available to evaluate semantic stabilization. Only few methods are yet suited for narrow folksonomies, where items are tagged only by the uploading user. Lin et al. [28] presents the Macro Tag Growth Method (MaTGM) that measures social vocabulary growth at a systemic level, looking at the social tagging system as a whole. In our setting, we consider each IBL group as an isolated social tagging system and thus, we apply MaTGM to compare the tag growth within these systems.

To that end, we first select the class that generated the most extensive tag data sets for both conditions (personal and collective) as representative groups. The selected data set is described in Table 2 as study TG. Then, for each group, we sort the posts (tag assignments) according to their timestamps, ending with the most recent item annotation. The tag growth after each post, is calculated as a value pair $(tg_i, f(tg_i))$, where tg_i is the cumulative number of tags, and $f(tg_i)$ is the cumulative number of unique tags occurring in i posts.

3.2 Recommender Accuracy

We evaluated the performance of the tag recommendation algorithms MP, BLL and Minerva utilizing the performance metrics recall, precision and f-measure, which are commonly used in recommender system research [31]. When calculating recall and precision, for each post, we determine the relation of tags recommended $\hat{T}_{u,r}$ to a user u for a resource r to the tags that the user assigned to a resource $T_{u,r}$.
Recall (R) indicates how well the recommendation supported the user, giving the relation between correctly recommended tags (i.e. the subset of recommended tags, that the user assigned to the resource) and the set of tags the user needed to describe the resource.

$$R(T_{u,r}, \hat{T}_{u,r}) = \frac{|T_{u,r} \cap \hat{T}_{u,r}|}{|T_{u,r}|} \tag{1}$$

Precision (P) is the proportion of tags that have been recommended correctly.

$$P(T_{u,r}, \hat{T}_{u,r}) = \frac{|T_{u,r} \cap \hat{T}_{u,r}|}{|\hat{T}_{u,r}|} \tag{2}$$

F-measure (F) combines recall and precision to their harmonic mean.

$$F = 2 \cdot \frac{(precision \cdot recall)}{(precision + recall)} \tag{3}$$

All metrics are averaged over the number of considered posts.

4. ALGORITHMS

The applied recommendation mechanisms have been extensively investigated in offline experiments [24, 22] where they showed promising results when applied on social bookmarking and TEL data sets. Notably, the cognitively inspired mechanisms consistently outperformed state-of-the-art tag recommendation algorithms such as Collaborative Filtering, FolkRank and even graph based methods.

Table 2: $|P|$ depicts the number of posts, $|U|$ the number of users, $|T|$ the number of tags, $|T_{unq}|$ the number of unique tags, $|AT_u|$ the average number of tags per user, $|AP_u|$ the average number of posts per user. *Vocabulary* refers to the data the tag recommendations were based on i.e., (P)ersonal or (C)ollective.

Research Topic	Vocabulary	$\lvert P\rvert$	$\lvert U\rvert$	$\lvert T\rvert$	$\lvert T_{unq}\rvert$	$\lvert AT_u\rvert$	$\lvert AP_u\rvert$
Soil ecosystems	P	9	6	17	11	2.3	1.5
	C	98	13	177	32	5.4	7.5
Biodiversity in cities	P	8	4	19	9	2.3	2.0
	C	35	14	75	24	3.9	2.5
Renewable resources	P	6	5	29	22	4.6	1.2
	C	12	8	34	19	4.1	1.5
Climate change	P	65	6	232	85	16.8	10.8
	C	83	10	297	86	16.4	8.3

(a) Properties of the preprocessed data sets extracted from eight inquiry groups.

Aspect	Vocabulary	$\lvert P\rvert$	$\lvert U\rvert$	$\lvert T\rvert$	$\lvert T_{unq}\rvert$	$\lvert AT_u\rvert$	$\lvert AP_u\rvert$
TG	C	83	10	297	86	16.4	8.3
	P	65	6	232	85	16.8	10.8
RA	C	228	38	584	153	15.4	6.0
	P	88	18	297	121	16.5	4.9

(b) Properties of the data sets taken into account for the investigations of two aspects: Tag growth (TG) and recommendation accuracy (RA).

4.1 Most Popular Tags

The Most Popular approach (MP) is a simple mechanism to rank tags according to their frequency of occurrence [19]. The algorithm is used as a baseline.

4.2 Base Level Learning Equation

Tagging resources on the web can be understood as a very basic form of communication, where people quickly retrieve word forms from their long-term memory [16], in order to provide textual labels for organizing their resources. In [23], we discuss and evaluate a personalized tag recommendation mechanism that mimics retrieval from human memory. The mechanism implements equations developed within the ACT-R architecture [1], in particular, to model the activation A_i and hence availability of elements in a person's declarative memory. Equation 4 comprises the base-level activation BLL and an associative component that represents semantic context. To model the semantic context, we look at the tags other users have assigned to the given resource, with W_j representing the frequency of appearance of a tag_j and with S_{ji} representing the normalized co-occurrence of tag_i and tag_j, as an estimate of the tags' strength of association.

$$A_i = BLL + \sum_j W_j S_{ji} \tag{4}$$

With equation 5, we estimate how useful an item (tag) has been for an individual person in the past, with n determining the frequency of tag use in the past, and t_j standing for recency, i.e. the time since a tag has been used for the j^{th} time. The parameter d models the power law function of forgetting and is in line with Anderson et al. [1] set to 0.5.

$$BLL = ln(\sum_{j=1}^{n} t_j^{-d}) \tag{5}$$

For the purpose of this study and taking into account data provided by the evaluation environment, the associative component cannot be calculated, as this component is based on tags other users have assigned to the very same content or item. weSPOT, however, is a narrow folksonomy (such as for instance Flickr), where content is generated and tagged only by one user. We thus make the assumption that BLL is the most accurate approach for our data set.

The most frequent tags of the user's inquiry group are considered, however, in order to continue collecting context information (i.e. tags that are new to a user). This is implemented in an additional recommendation approach denoted by BLL_U+ MP_G.

4.3 Minerva

The *Minerva* model aims to mimic a process of human categorization as introduced and described in Seitlinger et al. [36]. It consists of a simple network model with an input, a hidden and an output layer. The input layer is a vector P of n features that describe the item to be tagged. Within this study students assigned semantic features to their resources, by selecting suitable attributes from a drop-down menu. These attributes were drawn from the learning group's domain model which was provided by the teacher and describes the learning topic of a group in form of a formal concept lattice. For further information on the domain model please see Bedek et al. [4].

The hidden layer stores feature vectors of all data set items in a matrix S, such that S_{ik} is the activation of feature k in item i. Furthermore, in a tag matrix A, each data set item is associated with a tag vector A_i. Specifically, a tag activation value a_{ij} is defined as 1 if tag j was present in item i, and 0 otherwise. Taken the input vector as stimuli, the activation of single tags can be calculated. To this end we first compute the cosine similarity for the input feature vector P with each feature vector S_i in our matrix, following equation 6:

$$Sim_i = \frac{\sum_{k=1}^{n}(P_k \cdot S_{ik})}{\sqrt{\sum_{k=1}^{n} P_k^2} \cdot \sqrt{\sum_{k=1}^{n} S_{ik}^2}} \quad (6)$$

where P_k and S_{ik} are components of vector P and S_i respectively. Finally, we calculate the activation value t_j^{out} of tag j as the weighted sum of tag activation values over all items in the data set.

$$t_j^{out} = \sum_i Sim_i \cdot a_{ij} \quad (7)$$

The output layer is a ranked list of tags, with a maximum of five suggestions.

5. RESULTS AND DISCUSSION

This section presents the results of our evaluation. We evaluate the suitability of the algorithms described earlier for supporting learners' tagging processes. In line with the study design, all algorithms are applied in two modes: Personal (P), where the recommendation strategy draws on a single user's posting history, and collective (C), where the recommendation strategy draws on the prior posts of an entire group.

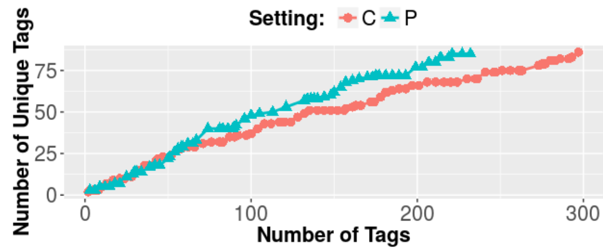
5.1 The Impact of Individual and Collaborative Tag Recommendations on Semantic Stabilization

The two plots illustrated in Figure 3 present the development of the tag vocabulary on a group level as described in Section 3.1. The graphs put side by side, the tag growth occurring in the collective group vocabulary condition (C) with the tag growth happening in the personal vocabulary condition (P), where students received their tag recommendations either based on collective tag traces or on personal tag traces, respectively. Figure 3a depicts the tag growth function according to the Macro Tag Growth Method and shows that while initially the vocabulary growth overlaps in both groups, group C starts to introduce less new vocabulary in relation to tags than group P. In other words, we can observe that students in the collective condition start to pick up the vocabulary of their peers faster. This result is even stronger when considering that a greater number of users contributed to the tagging data of the collective condition than to the data of the personal vocabulary condition (see Table 2). This indicates a positive effect of collective tag recommendations on semantic stabilization. Figure 3a provides additional insights into the timing of the process. We can observe that the two tag growth functions clearly diverge after about 40 added posts.

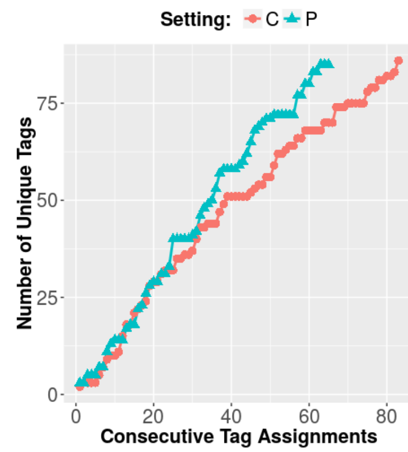
5.2 The Accuracy of Cognitive-Inspired Tag Recommendation Strategies in an Online Data Setting

This section presents the results of our evaluation study in respect to recommendation accuracy. Table 3 provides the number of observations (see column N_T) and accuracy estimates (R, P and F) for each recommender.

The table discloses the impact of the two variables algorithm and data set on performance: BLL appears to reach higher estimates than *Minerva* (relative to *MP*) under the personal vocabulary condition, with the opposite being true for the collective condition.



(a) Tag growth function according to the Macro Tag Growth Method.

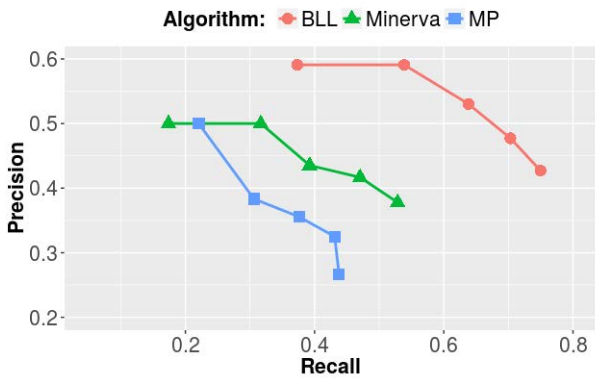


(b) Number of unique tags accumulated with consecutive tag assignments.

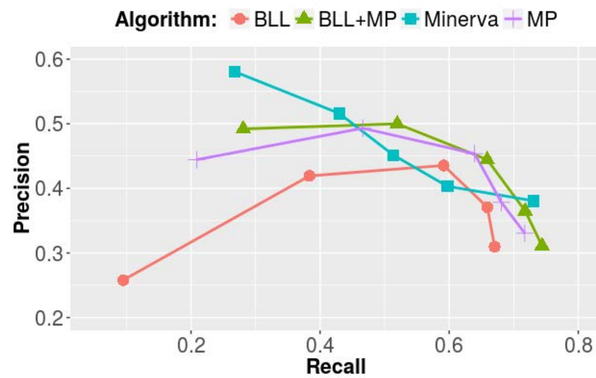
Figure 3: The plots show the development of tagging vocabulary on a system (inquiry-based learning group) level. The two line graphs depict the between-subject variables of the study, that distinguish between the settings: collective (C) and personal (P).

Table 3: Properties of the analysed data set, structured by the applied algorithm. Data defines whether the algorithm was calculated on a user's personal word trace P, an inquiry groups collective word traces C or a *Mixed* approach PC considering both type of data. N_T depicts the number of tagged resources, we derived from the online evaluation. The metrics recall, precision and f-measure are mean values and standard deviations of $R@5$, $P@5$ and $F@5$, respectively.

	Algorithm	N_T	$P@5$	$R@5$	$F@5$
P	*MP*	30	0.26 (0.25)	0.44 (0.36)	0.31 (0.27)
	Minerva	36	0.38 (0.32)	0.53 (0.39)	0.41 (0.31)
	BLL	22	**0.43** (0.28)	**0.75** (0.33)	**0.50** (0.26)
C	*MP*	72	0.33 (0.21)	0.72 (0.36)	0.42 (0.23)
	BLL	62	0.31 (0.23)	0.67 (0.37)	0.39 (0.23)
	Minerva	31	0.38 (0.28)	0.73 (0.38)	0.46 (0.30)
PC	$BLL + MP$	63	0.31 (0.21)	0.74 (0.38)	0.41 (0.25)



(a) Personal vocabulary condition: tag recommendations are based on a user's personal tagging traces.



(b) Collective vocabulary condition: tag recommendations are based on the learning group's collective tagging traces.

Figure 4: Recall/Precision plots illustrating the accuracy of recommendation algorithms in the personal and the collective vocabulary condition. BLL applied in the personal setting performs best over all considered recommendation approaches. In the collective condition, best results can be achieved for BLL+MP and Minerva.

In line with this descriptive pattern, a 2 (Personal vs. Collective) × 3 (MP vs. BLL vs. Minerva) ANOVA on F reveals no significant main effects - either for vocabulary, $F(1, 44)$=1.22, $n.s.$, nor algorithm, $F(2, 44)$=2.35, $n.s.$ - but a significant interaction between these two factors, $F(2, 44) = 4.33$, $p < .05$.

Results indicate that in the personal setting the BLL approach, which mimics the activation of words in a person's memory as a function of frequency and recency, performs best. On the other hand BLL applied in the collective vocabulary condition performed very poorly (see also Figure 4).

Also, we can see that the recommender $Minerva$ showed better performance in the collective, than in the personal vocabulary condition. While a model that categorizes according to semantic context should be able to depict both, personal and collective data, it is fair to assume that the size of the data set plays a crucial role. We believe the approach will become more accurate with the growing extent of the data set. Hence, we draw two conclusions. Firstly, $Minerva$ performs better on collective than on personal tagging traces, as the data set is likely to be more extensive. Secondly, the performance of the algorithm will enhance with the time of use.

This corroborates our expectations, as we can assume that student's interests within a group differ but are individually relatively stable within the short period of a school project. The individual developments of the students within a topic can be further depicted with the introduction of recency, as implemented in BLL_U.

A very interesting result constitutes the moderate performance of BLL_U+ MP_G, since it is contrary to results from offline TEL data studies such as presented in Kopeinik et al. [22] where the approach clearly outperforms remaining recommendation strategies. This conforms with the assumed task difference between online and offline studies, according to which we either evaluate an support or a prediction task, respectively.

5.3 Data Sets

If we look at Table 2, we can observe that the tagging frequency varies greatly among the groups. Students that participated in the study used the environment in the course of biology lessons. However, the IBL project work did not contribute to their marking. Also, they were encouraged to tag, but there was no particular monitoring of this process taking place. Thus, some groups showed more motivation and participated more actively in the projects and within the environment than others.

Another aspect is that there is significantly less data available for vocabulary condition P, where users tag recommendations were based on their individual tagging history. Due to the cold start problem, students in condition P had no initial tag seeds provided but rather had to come up with their own personal tag traces to initiate the tag recommendation process. We believe the resulting lack of tagging support, played a crucial role when students did not tag their contributions or tagged their contributions in unusual ways (see Section 2.4). This is in line with previous findings (e.g., [26]) that underline the need for support students have in the tagging process.

6. CONCLUSION

This paper presents a real-world evaluation investigating the application of tag recommender approaches from two perspectives: First, by dividing students in two groups receiving either tag recommendations based on personal or collective tag traces, we gain insights into the effect of collective tag recommendations on the semantic stabilization process of collective learning groups.

Second, we evaluate the performance of two tag recommender approaches that imitate human behaviour, in particular the process of human categorization and the retrieval of words from memory. The algorithms, Minerva and base-level learning equation (BLL), as well as MP as a baseline, were applied as within-subject variables, either on the basis of the collective or the personal tagging history.

Our results demonstrate that selecting recommendations from the collective vocabulary, i.e., exposing a learner to others' tags, is much more effective to promote semantic stabilization than drawing from the personal vocabulary and thus, displaying only individual tags. Furthermore, the results suggest that searching for relevant tags in the collective's vocabulary benefits strongly from considering usage frequency and semantic context, i.e., from a strategy implemented by Minerva. The information of recency, on the other hand, appears to show advantages when aiming to identify relevant tags within the personal vocabulary.

One practical design implication is thus that semantic stabilization within the setting of inquiry-based group learning can be supported well by recommenders that both draw on data of the whole collective and are sensitive to the semantic context of learners' search results in order to estimate tag choice probabilities. In case of an individual learning setting, however, we suggest applying recommenders that focus on information about time and frequency of past tag choices to predict their current availability in a learner's memory and hence, relevance for the current learning episode.

We are aware of the limited evaluation data which is a consequence of the selected real-world learning setting. Data in such learning environments is typically sparse, which has also been the reason for restricting the experiment to the three algorithm set-up. By contrast, results from offline data studies can compare a multitude of options. However, we argue that those results are limited in their reliability, as unlike real-world studies, offline data do not allow for investigation of recommendation strategies' ability to support users in their tasks, but solely evaluate the prediction of a user's behaviour.

In future work we are planning to strengthen our argument by introducing students' learning as an additional dependent variable. This will allow for further investigation of the correlation between individual's learning progress and the development of a common terminology in their learning group.

7. ACKNOWLEDGMENTS

This work is supported by the EU funded projects weSPOT (Grant Agreement 318499), Learning Layers (Grant Agreement: 318209) and AFEL (Grant Agreement: 687916), the Austrian Science Fund (FWF) Projects MERITS (Grant No P25593-G22) and OMFix (Grant No P27709-G22), and the Know-Center. The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. We are also very grateful to Verein für Bildung und Erziehung der Grazer Schulschwestern and particularly to the teacher Jürgen Mack for the great support in implementing the study in his lessons.

8. REFERENCES

[1] J. R. Anderson and L. J. Schooler. Reflections of the environment in memory. *Psychological science*, 2(6):396–408, 1991.

[2] M. Anjorin, I. Dackiewicz, A. Fernández, C. Rensing, et al. A framework for cross-platform graph-based recommendations for tel. In *Proceedings of the 2nd workshop on recommender systems in technology enhanced learning*, pages 83–88, 2012.

[3] S. Bateman, C. Brooks, G. McCalla, and P. Brusilovsky. Applying collaborative tagging to e-learning. *Proceedings of ACM WWW*, 3(4), 2007.

[4] M. A. Bedek, S. Kopeinik, B. Prunster, and D. Albert. Applying the formal concept analysis to introduce guidance in an inquiry-based learning environment. In *Advanced Learning Technologies (ICALT), 2015 IEEE 15th International Conference on*, pages 285–289. IEEE, 2015.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl. Taxonomy of educational objectives, handbook 1: The cognitive domain, 1956.

[7] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs. A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5-6):318–331, 2012.

[8] E. Diaz-Aviles, M. Fisichella, R. Kawase, W. Nejdl, and A. Stewart. Unsupervised auto-tagging for learning object enrichment. In *Towards Ubiquitous Learning*, pages 83–96. Springer, 2011.

[9] H. Drachsler, H. Hummel, and R. Koper. Recommendations for learners are different: Applying memory-based recommender system techniques to lifelong learning. In *TENC: Publications and Preprints. Paper presented at the SIRTEL workshop of EC-TEL 2007*, volume 1. Keur der Wetenschap, 2007.

[10] H. Drachsler, K. Verbert, O. C. Santos, and N. Manouselis. Panorama of recommender systems to support learning. In *Recommender systems handbook*, pages 421–451. Springer, 2015.

[11] E. Duval. Attention please!: learning analytics for visualization and recommendation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, pages 9–17. ACM, 2011.

[12] F. Font, J. Serrà, and X. Serra. Analysis of the impact of a tag recommendation system in a real-world folksonomy. *ACM Trans. Intell. Syst. Technol.*, 7(1):6:1–6:27, Sept. 2015.

[13] W.-T. Fu. The microstructures of social tagging: a rational model. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 229–238. ACM, 2008.

[14] W.-T. Fu and W. Dong. Collaborative indexing and knowledge exploration: A social learning model. *IEEE Intelligent Systems*, 27(1):39–46, 2012.

[15] W.-T. Fu, T. G. Kannampallil, and R. Kang. A semantic imitation model of social tag choices. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 66–73. IEEE, 2009.

[16] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, pages 211–220. ACM, 2007.

[17] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *Proceedings of the 31st annual international ACM SIGIR conference on*

Research and development in information retrieval, pages 531–538. ACM, 2008.

[18] D. L. Hintzman. Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, 16(2):96–101, 1984.

[19] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007*, pages 506–514. Springer, 2007.

[20] M. K. Khribi, M. Jemni, and O. Nasraoui. Recommendation systems for personalized technology-enhanced learning. In *Ubiquitous learning environments and technologies*, pages 159–180. Springer, 2015.

[21] A. Klašnja-Milićević, M. Ivanović, and A. Nanopoulos. Recommender systems in e-learning environments: a survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, 44(4):571–604, 2015.

[22] S. Kopeinik, D. Kowald, and E. Lex. Which algorithms suit which learning environments? a comparative study of recommender systems in tel. In *European Conference on Technology Enhanced Learning*, pages 124–138. Springer, 2016.

[23] D. Kowald, S. Kopeinik, P. Seitlinger, T. Ley, D. Albert, and C. Trattner. Refining frequency-based tag reuse predictions by means of time and semantic context. In *Mining, Modeling, and Recommending'Things' in Social Media*, pages 55–74. Springer, 2015.

[24] D. Kowald, P. Seitlinger, S. Kopeinik, T. Ley, and C. Trattner. Forgetting the words but remembering the meaning: Modeling forgetting in a verbal and semantic tag recommender. In *Mining, Modeling, and Recommending'Things' in Social Media*, pages 75–95. Springer, 2015.

[25] A. Kuhn, C. Cahill, C. Quintana, and S. Schmoll. Using tags to encourage reflection and annotation on data during nomadic inquiry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 667–670. ACM, 2011.

[26] A. Kuhn, B. McNally, S. Schmoll, C. Cahill, W.-T. Lo, C. Quintana, and I. Delen. How students find, evaluate and utilize peer-collected annotated multimedia data in science inquiry with zydeco. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3061–3070. ACM, 2012.

[27] T. Ley and P. Seitlinger. Dynamics of human categorization in a collaborative tagging system: How social processes of semantic stabilization shape individual sensemaking. *Computers in human behavior*, 51:140–151, 2015.

[28] N. Lin, D. Li, Y. Ding, B. He, Z. Qin, J. Tang, J. Li, and T. Dong. The dynamic features of delicious, flickr, and youtube. *Journal of the American Society for Information Science and Technology*, 63(1):139–162, 2012.

[29] G. Macgregor and E. McCulloch. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library review*, 55(5):291–300, 2006.

[30] N. Manouselis, H. Drachsler, R. Vuorikari, H. Hummel, and R. Koper. Recommender systems in technology enhanced learning. In *Recommender systems handbook*, pages 387–415. Springer, 2011.

[31] L. B. Marinho, A. Hotho, R. Jäschke, A. Nanopoulos, S. Rendle, L. Schmidt-Thieme, G. Stumme, and P. Symeonidis. *Recommender systems for social tagging systems*. Springer Science & Business Media, 2012.

[32] A. Mathes. Folksonomies-cooperative classification and communication through shared metadata, 2004. Available at: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[33] K. Niemann. Automatic tagging of learning objects based on their usage in web portals. In *Design for Teaching and Learning in a Networked World*, pages 240–253. Springer, 2015.

[34] A. Protopsaltis, P. Seitlinger, F. Chaimala, O. Firssova, S. Hetzner, K. Kikis-Papadakis, and P. Boytchev. Working environment with social and personal open tools for inquiry based learning: Pedagogic and diagnostic frameworks. *The International Journal of Science, Mathematics and Technology Learning*, 20(4):51–63, 2014.

[35] R. Raffelsiefen. Semantic stability in derivationally related words. *Amsterdam Studies in the Theory and History of Linguistic Science*, 4:247–268, 1998.

[36] P. Seitlinger, T. Ley, and D. Albert. An implicit-semantic tag recommendation mechanism for socio-semantic learning systems. In *Open and Social Technologies for Networked Learning*, pages 41–46. Springer, 2013.

[37] M. Specht, M. Bedek, E. Duval, P. Held, A. Okada, K. Stevanov, E. Parodi, K. Kikis-Papadakis, and V. Strahovnik. Wespot: Inquiry based learning meets learning analytics. In *3rd international conference on e-Learning*, pages 15–20, 2012.

[38] C. Trattner, D. Kowald, P. Seitlinger, S. Kopeinik, and T. Ley. Modeling activation processes in human memory to predict the use of tags in social bookmarking systems. *The Journal of Web Science*, 2(1):1–16, 2015.

[39] K. Verbert, H. Drachsler, N. Manouselis, M. Wolpers, R. Vuorikari, and E. Duval. Dataset-driven research for improving recommender systems for learning. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, pages 44–53. ACM, 2011.

[40] C. Wagner, P. Singer, M. Strohmaier, and B. Huberman. Semantic stability and implicit consensus in social tagging streams. *IEEE Transactions on Computational Social Systems*, 1(1):108–120, 2014.

[41] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Collaborative web tagging workshop at WWW2006, Edinburgh, Scotland*, 2006.