From Task Classification Towards Similarity Measures for Recommendation in Crowdsourcing Systems

Steffen Schnitzer and **Svenja Neitzel** and **Christoph Rensing**
Multimedia Communications Lab - Technische Universität Darmstadt, Germany
steffen.schnitzer@kom.tu-darmstadt.de
svenja.neitzel@kom.tu-darmstadt.de
christoph.rensing@kom.tu-darmstadt.de

Abstract

Task selection in micro-task markets can be supported by recommender systems to help individuals to find appropriate tasks. Previous work showed that for the selection process of a micro-task the semantic aspects, such as the *required action* and the *comprehensibility*, are rated more important than factual aspects, such as the payment or the required completion time. This work gives a foundation to create such similarity measures. Therefore, we show that an automatic classification based on task descriptions is possible. Additionally, we propose similarity measures to cluster micro-tasks according to semantic aspects.

1 Introduction and Related Work

A preceding user study (Schnitzer et al. 2015) shows that the similarity of tasks is an important factor for workers when selecting tasks in crowdsourcing platforms. Another preceding study (Schnitzer et al. 2016) identified the most important similarity aspects for workers. The semantic aspects, in contrast to the factual aspects, were found to be the five most highly rated similarity aspects, with *required action* and *comprehensibility* coming first and second.
This work provides a foundation to leverage such semantic aspects for recommending tasks in crowdsourcing platforms using similarity measures based on task descriptions. To show, that task descriptions in micro-task markets are diverse and informative enough to support a recommendation, our first approach classifies tasks into predefined categories. Therefore, an automatic classification of tasks is implemented and evaluated on a dataset of 1466 micro-tasks retrieved from the platform *Microworkers*. We compare different classification approaches by evaluating different feature sets and their combinations as well as several classification algorithms. This allows us to conclude, that the employed methods are capable of classifying micro-tasks into logical categories. On this basis, we conclude that similarity measures for the identified semantic similarity aspects can be created from task descriptions. Therefore, we propose a first idea for creating similarity measures based on task descriptions considering the semantic aspects *required action* and *comprehensibility*.

One approach to task recommendation has been proposed by Yuen, King, and Leung (2012), considering task properties, worker performance and history of the worker's completed tasks. The two methods proposed and compared by Ambati, Vogel, and Carbonell (2011) use a classification as well as an approach based on semantic similarities. Mavridis, Gross-Amblard, and Miklós (2016) apply a taxonomy based skill modeling approach to optimize task assignment quality. Within our classification approach, the textual information from the micro-tasks is used to classify them into the categories provided by the platform. Arora, Ganguly, and Jones (2015) present an approach for classifying questions posted to a Q&A platform. In a very similar domain to micro-tasks, Schnitzer et al. (2014) and Schmidt, Schnitzer, and Rensing (2016) use a *tf-idf* based approach and an ensemble classifier in order to classify job offers.

2 Task Classification

Dataset and Preprocessing The dataset of 1466 micro-tasks was gathered between October and December 2015 from the micro-task market platform *Microworkers*. For each task we extract the ID, title, description, proof, category, employer, payment, time to finish, time to rate, no. of jobs available/done, success rate and countries the task is available in. A number of common preprocessing steps are applied to the textual task attributes and some meta information about the original text given by the HTML structure is extracted and stored as additional attributes.

Classification of Micro-Tasks For classification, the machine learning tool Weka is used. To identify the most accurate setup for classification, four different feature sets (see Table 1) are extracted and six different classifiers are trained on every combination of the feature sets. The Weka implementations of six different classifiers are used to evaluate the performances of Naive Bayes, Random Forest, K Nearest Neighbors (*IBk*), Support Vector Machine (*SMO*), a rule based classifier (*JRip*) and a decision tree (*J48*).

Evaluation A 10-fold stratified cross-validation is executed on the dataset for each classifier using each feature set. The results obtained for the three best performing classifiers (JRip, SMO and Random Forest) in terms of weighted average F1-score are given in Table 2. The content feature set using a tf-idf approach achieves the best results over all

Table 1: Feature sets.

Feature Set	Features
factual	payment, time to rate, time to finish, positions, payment per minute, employer, countries
content	n-grams
structural	word count, no. of bullet points, avg. words per sentence, avg. commas per sentence, avg. chars per word, avg. paragraph length, avg. line length, readability (Gunning Fog Index (Gunning 1952)), lexical diversity (Dickinson et al. 2015)
semantic	URL hosts, named entities, sentiment

Table 2: F1-scores for different classifiers and feature sets.

Feature Set	Random Forest	JRip	SMO
factual	0.86	0.82	0.73
structural	0.81	0.74	0.54
semantic	0.83	0.75	0.84
content (tf-idf)	**0.92**	**0.92**	**0.94**

classifiers. The SMO classifier obtains the highest F1-score of 0.94 using the content feature set. However, it is outperformed by the two other classifiers when using the factual or structural feature set. Random Forest turns out to be the most stable classifier across the four feature sets.

This evaluation shows in general, that it is possible to reproduce the task categories and that a classification of micro-tasks is feasible. Content features were shown to be the best performing feature set, while the SMO classifier provided the best results among the classifiers. A per class evaluation showed further, that all classes with at least 10 examples can be classified with an F1-score above 0.7.

3 Similarity Measures for Micro-Tasks

Task similarities based on the semantic aspects *required action* and *comprehensibility*, that were found to be relevant in the preceding study (Schnitzer et al. 2016), cannot be produced by a classification approach. The classification considers binary category membership, while similarities rely on continuous measures. However, the insights about the applicability of certain features for the classification task can be used and extended to propose an approach for calculating task similarities based on these aspects. As there is no labeled data and no predefined classes for *required action* and *comprehensibility* for micro-tasks, an unsupervised approach is necessary. In the following we propose certain features for the similarity measures that are specific for each of the semantic aspects.

Features for *required action* To measure how similar two tasks are in their *required action*, we consider verb phrases within the task descriptions. The verb phrases in the task description are chosen, as we expect them to reflect the actions that are required to solve the task. Two task descriptions that share some verb phrases are likely to be similar regarding their *required action*. However, many verb phrases bear similar meaning, even though the vocabulary is not exactly the same. Therefore, we apply word similarities from *WordNet* (Abdalgader and Skabar 2010), (Chang, Lee, and Wang 2016).

Features for *comprehensibility* To measure similarities in comprehensibility, we adopt features from the structural and content feature set used in the classification approach. Those features are: word count, number of bullet points, average words per sentence, average commas per sentence, average chars per word, average paragraph length, average line length, readability and lexical diversity. Additionally, we add the *ratio of unusual words* in the tasks' descriptions, which is computed as the percentage of the words in the task's description that are neither contained in more than five tasks in the whole corpus nor in the English word list obtained from a Unix operating system.

Evaluation First experiments that apply the similarity measures to cluster the tasks in the dataset show good results. The category distribution of clusters regarding *required action* (see Table 3) shows, that some clusters seem to model known categories while others include tasks from many different categories.

Table 3: Category distribution for selected clusters regarding *required action*.

Category	$\mathbf{A}_1$	$\mathbf{A}_6$	$\mathbf{A}_{11}$
Blog/Website Owners	-	-	0.11
Facebook	0.08	-	-
Google	0.10	-	-
Mobile Applications	0.02	-	0.89
Other	0.16	-	-
Promotion	-	0.03	-
Search, Click, Engage	0.47	-	-
Sign up	0.09	0.94	-
Youtube/Vimeo/...	0.04	-	-
Various	0.04	0.03	-

4 Conclusion

This paper shows how the content of task descriptions can be used to create a classification of micro-tasks. It also proposes two additional similarity measures for micro-tasks. The evaluation shows that a classification is feasible using the proposed setup. We also propose similarity measures that can be applied to find similarities between micro-tasks. The proposed similarity measures can model similarities, that are different from the known categories.

References

Abdalgader, K., and Skabar, A. 2010. Short-text similarity measurement using word sense disambiguation and synonym expansion. In *Australasian Joint Conference on Artificial Intelligence*, 435–444. Springer.

Ambati, V.; Vogel, S.; and Carbonell, J. G. 2011. Towards Task Recommendation in Micro-Task Markets. In *Proceedings of the 3rd Human Computation Workshop (HCOMP)*.

Arora, P.; Ganguly, D.; and Jones, G. J. F. 2015. The Good, the Bad and Their Kins: Identifying Questions with Negative Scores in StackOverflow. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, 1232–1239. ACM.

Chang, J. W.; Lee, M. C.; and Wang, T. I. 2016. Integrating a semantic-based retrieval agent into case-based reasoning systems: A case study of an online bookstore. *Computers in Industry* 78:29–42.

Dickinson, T.; Fernandez, M.; Thomas, L. A.; Mulholland, P.; Briggs, P.; and Alani, H. 2015. Identifying Prominent Life Events on Twitter. In *Proceedings of the 8th International Conference on Knowledge Capture*. ACM.

Gunning, R. 1952. *The Technique of Clear Writing*. McGraw-Hill, New York.

Mavridis, P.; Gross-Amblard, D.; and Miklós, Z. 2016. Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing. In *Proceedings of the 25th International Conference on World Wide Web*, 843–853.

Schmidt, S.; Schnitzer, S.; and Rensing, C. 2016. Text classification based filters for a domain-specific search engine. *Computers in Industry*.

Schnitzer, S.; Schmidt, S.; Rensing, C.; and Harriehausen-Mühlbauer, B. 2014. Combining active and ensemble learning for efficient classification of web documents. *Polibits* 49:39–45.

Schnitzer, S.; Rensing, C.; Schmidt, S.; Borchert, K.; Hirth, M.; and Tran-Gia, P. 2015. Demands on Task Recommendation in Crowdsourcing Platforms - The Worker's Perspective. In *Proceedings of the CrowdRec Workshop at ACM Recommender Systems Conference*.

Schnitzer, S.; Neitzel, S.; Schmidt, S.; and Rensing, C. 2016. Perceived Task Similarities for Task Recommendation in Crowdsourcing Systems. In *Proceedings of the 25th International Conference Companion on World Wide Web*.

Yuen, M.-C.; King, I.; and Leung, K.-S. 2012. Task Recommendation in Crowdsourcing Systems. In *Proceedings of the 1st International Workshop on Crowdsourcing and Data Mining*. ACM.