Determining the Outdatedness Level of Knowledge in Collaboration Spaces Using A Machine Learning-Based Approach

Zahra Grosser[1], Andreas Schmidt[1], Martin Bachl[1], and Christine Kunzmann[2]

[1] Karlsruhe University of Applied Sciences, Karlsruhe, Germany
[2] Pontydysgu Ltd., Germany

Abstract. Knowledge development increasingly takes place in shared collaborative document spaces where the number of created documents and participating individuals is growing continuously. Over time, such spaces leave members wondering about the status and reliability of documents, e.g., whether it is work-in-progress, abandoned, or outdated. In this paper, we investigate the outdatedness of knowledge by implementing an automated approach that enables individuals to become aware of the outdatedness level of knowledge in documents. More specifically, we address this challenge for a wiki-based collaboration space based on MediaWiki, in which documents are represented as wiki articles. In our proposed model, an automatic classification technique based on machine learning approaches is adopted to assess the outdatedness level of wiki articles. Our model evaluates three feature categories including article-based, contributor authoritativeness, and category-dependent features. Through different experiments, we analyse how different feature sets contribute to the overall classification performance. We also show that assessing the outdatedness of articles within specific categories can significantly improve the classification performance. Our experimental results demonstrate that our model achieves a promising performance. The proposed model and workflow is applicable to other collaboration spaces and other indicators such as the relevancy of documents.

Keywords: collaboration spaces, wiki, outdatedness, knowledge maturing, machine learning, topic categories, classification

1 Introduction

Agile work needs collaboration spaces that facilitate self-organization. Collaboration spaces provide an inter-connected environment for individuals to work and interact with each other. Individuals communicate through joint activity and they create, share, and exchange knowledge which is represented by artefacts, such as documents. In a large-scale collaboration space, where multiple communities of individuals work together, a document can be shared and modified by everybody. Moreover, there are no mechanisms that help deciding whether a given document can be relied upon, whether it is controversial, work-in-progress, or

even outdated. Problem can arise by using such documents particularly for new and peripheral members. The *Knowledge Maturing Model* proposed by Maier & Schmidt [15] offers a developmental perspective on collective knowledge by identifying different phases and corresponding characteristics of knowledge maturing. Indicators for such phases can be of great help for enhancing collaborative spaces with visual cues or filtering mechanisms. One particular research area of knowledge maturing is to capture the phenomenon of outdated knowledge, also referred to as outdatedness. Therefore, we focus our research activity on using indicators to explore the outdatedness of knowledge. Uncertainty about whether knowledge in collaboration spaces is up-to-date, constitutes a major barrier, especially when those spaces become larger in terms of number of documents, number of individuals, and usage time. Although an alternate approach to deal with this problem is regular review of documents, the marking of content for periodic review, as well as performing the actual review tasks in larger networks requires considerable effort. Therefore, automated outdatedness indicators are necessary and appear promising by creating awareness when sharing and using documents in such collaboration spaces. From the design point of view we are interested to employ visual cues as indicators.

Following this idea, the main contribution of this paper is to propose a model and workflow to automatically determine the outdatedness level of collective knowledge in collaboration spaces. There are two distinct concepts associated to the phenomenon of outdatedness of knowledge. First, knowledge can become irrelevant in a specific context, for example when the focus and the directions are changing for a professional community. Second, knowledge can become invalid with new evidence, for example the documents no longer reflect the current status of a conversation. In this paper, we cover mainly invalidity.

In our experiments, we focus on an active collaboration space of a research project with over 600 articles, called the EmployID wiki. At the current state of our experiments, the project was active for 2.5 years and there are still 1.5 years to go. More specifically, the contributions of this paper are to:

- investigate the value of using a supervised machine learning technique, known as Support Vector Machines (SVM), to determine the outdatedness level of wiki articles through classification.
- evaluate the performance of the model using a number of features extracted from attributes of wiki articles including content, edit history, and contributor authoritativeness.
- introduce novel features, namely category-dependent features, with respect to specific characteristics of groups of articles in a same category by applying text analysis techniques.
- evaluate the classification performance by employing a feature selection method.

The remainder of this paper is structured as follows. Section 2 discusses related work. Section 3 introduces the EmployID wiki. Section 4 formally defines the problem of wiki article outdatedness classification. Section 5 describes our

proposed model and workflow and explains how we approach the wiki article outdatedness classification problem. Section 6 presents our experimental setup. Section 7 demonstrates our experimental results of the wiki article outdatedness classification. Section 8 concludes this paper and sketches some future work.

2 Related Work

This section briefly surveys the research areas related to our work, which are knowledge maturing, outdatedness determination, feature extraction, and classification using machine learning techniques. As a first study related to the underlying *Knowledge Maturing Model* [15], Braun & Schmidt [3] perform analysis of Wikipedia articles with respect to different article features to identify measures for maturity. In this work, they use the *excellent* article badge as a gold standard.

Regarding the outdatedness determination, to the best of our knowledge, there is no work dedicated for automatically determining the outdatedness of wiki articles. However, a number of approaches address automated assessment of the quality of User Generated Content (UGC). Quality assessment is a multidimensional concept, which comprises several criteria such as trustworthiness, accuracy, reliability, and relevancy assessment. In the following, we review a number of UGC quality assessment models that have been developed for social media platforms, including web forums and Wikipedia. These research works are related to our study as they employ machine learning techniques to train classification functions, which map sets of object features to sets of class attributes.

The literature provides numerous approaches to assess the quality of Wikipedia articles. All of these approaches rely on feature extraction. What differentiates the approaches from each other is the variety of feature sets and their complexity. Hu et al. [10] evaluate the effectiveness of combining contributor authoritativeness features with the character count feature to measure the quality of articles in Wikipedia, which results in performance improvements of their model. Both these features are also employed in our model.

Anderka et al. [2] provide a comprehensive overview of several features, organized along the four dimensions content, structure, network, and edit history to predict quality flaws in UGC. Additionally, they develop a flaw-specific view that combines feature selection, expert rules, and multilevel filtering to explore essential features for the prediction of a certain quality flaw. Our research is related to this work in that we employ feature selection techniques to find effective subsets of features with respect to specific categories of related articles for which multiple classifiers are trained in parallel.

There is also another study, which applies domain-specific features for quality evaluation of Wikipedia medical articles [7]. This work creates a model, which combines some baseline structural and linguistic features with specific features extracted from the medical domain by employing Natural Language Processing (NLP) techniques. This study reports that applying domain-specific features,

such as domain informativeness and category, can significantly improve the automatic classification results.

In a case study on web forums, Weimer et al. [20] is one of the first efforts. They present a binary SVM classifier to predict the quality of forum posts by using content and lexical features. They observed that the classification performance is dependent on specific topics discussed. Their model generates better performance on a dataset containing discussions about software when compared to another dataset containing content from miscellaneous topics. This is due to the general assumption, that by grouping related content within categories before classification, better classification performance is achieved. This idea is also adopted in our work. Wanas et al. [19] extend this work by proposing a model to classify the quality of forum posts into three quality classes (i.e., low, medium, and high) on a larger dataset. This work is different from previous works in that they evaluate edit history features, which are also represented in our study. The effectiveness of these features for forum posts quality classification is demonstrated in their work.

Chai et al. [4] propose a model that assesses post usage behavior of the forum community to measure the quality of forum posts. Regarding our research, usage-based features, such as view count or dwell time, don't seem promising. That's because an article could be opened several times randomly while its view count increases, and at the same time it does not reflect the popularity of the page. This work is further extended in [5] by adding content, reputation (contributor authoritativeness), and temporal features, which are employed by different classifiers. Furthermore, they conduct a number of feature preprocessing tasks such as normalization, discretization and feature selection. Their work performs a number of classification experiments. Their proposed model is validated on three forum datasets and existing models known from [19, 20] are outperformed in their experiments. Utilizing different classifiers and applying feature discretization techniques appear to be promising for future investigation.

3 The EmployID Wiki

The EmployID wiki is a wiki-based collaboration space based on MediaWiki with Semantic MediaWiki extensions. It is used by the EmployID research project[1] team for organizing knowledge and all project information related to workplace learning and technologies supporting public employment services. The EmployID wiki consists of roughly 50 project members and 600 content pages from about 2.5 years (out of a total of four years project duration) of project work. The wiki allows individuals to create and share knowledge, collaborate on projects, and manage projects more effectively. Articles of the EmployID wiki cover various aspects of the project, including project news, brainstorming and ideas, project reports, guidelines, to-do tracking, meeting minutes, and scrum team overview pages. In addition to the content, metadata has been added to articles, such

[1] www.employid.eu

as categories. Individuals categorize articles and files by adding one or more category tags as wiki markup. That makes it easy to browse related articles with same characteristics.

The data of the wiki pages, including the entire editing history, is stored in a database. The wiki database layout contains dozens of tables including pages and their content, individuals and their preferences, metadata, etc.

The EmployID wiki is a real-life and vibrant data source. It is unknown how many pages are up-to-date, valid, or relevant. For example, there are articles, which should be updated and there are other articles, which are not valid anymore, describing what was important in the past. That makes the EmployID wiki a valuable resource for the purpose of our research to learn how information is getting outdated over time. Therefore, providing indicators for project members enables them to stay informed about outdated information or to be aware of articles, which should be reviewed from time to time.

4 Wiki Article Outdatedness Classification Definition

The phenomenon of outdatedness of knowledge in a wiki-based collaboration space is defined as follows:

Definition 1. An outdated wiki article is an article whose content is no longer considered to be *true*, *good*, *useful*, or has become *invalid*[1].

That is, definition 1 is based on facts and evidence as opposed to the perceived outdatedness of wiki articles, which leads to definition 2.

Definition 2. An outdated wiki article is an article whose content is marked as outdated by the ratings of the majority of individuals.

Invalidity and outdatedness of wiki articles need to be introduced as concepts, so that wiki members become aware of them. Therefore, the main contribution of this paper is to provide a model and workflow that automatically determines the outdatedness of wiki articles based on features extracted from various aspects of wiki articles. The generated feature vector is then passed into a supervised machine learning algorithm that learns from article outdatedness ratings derived from the feedback of individuals.

The problem of determining the outdatedness of wiki articles is formally defined as a multi-class classification problem. We focus on the definition of the wiki article outdatedness classification, the wiki article structure, and the representation of the wiki article.

Let $A = \{a_1, a_2, \ldots, a_{|A|}\}$ be a set of wiki articles to be classified. Based on ratings of articles, we evaluate three article outdatedness classes, which can be defined as $C = \{c_1 = low, c_2 = medium, c_3 = high\}$. Each article is represented as a set of features we use for classifying the outdatedness of wiki articles as defined for a_i in (1):

$$a_i = \{f_1^i, f_2^i, \ldots, f_{|F|}^i\} \tag{1}$$

[1] http://results.learning-layers.eu/scenarios/knowledge-maturing/

Then, we define $F(a_i, c_k)$ be a decision matrix which represents whether a_i belongs to c_k where $k = \{1, 2, 3\}$ as defined in (2).

$$F(a_i, c_k) : A \times C \rightarrow \{True, False\} \tag{2}$$

The task of the wiki article outdatedness classification is to approximate this unknown function for all articles by means of a machine learning classifier over of training samples of the wiki article dataset.

5 Model Definition

The general workflow of the wiki article outdatedness determination is based on the standard workflow for classification problems, and in our case it consists of five major phases, namely data collection, data preprocessing, feature extraction, feature preprocessing, and classification using machine learning techniques, as depicted in Figure 1.

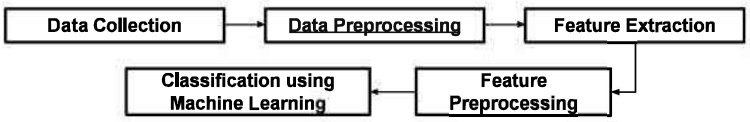

Fig. 1: The wiki article outdatedness determination model workflow

During the first phase large amounts of raw data are extracted from the EmployID wiki dataset such as MediaWiki common tables.

The data preprocessing phase refers to preparing and cleaning data before running the feature extraction. This is an important task, which can assist in the improving classification performance. Removing unnecessary and irrelevant information is an example of the data preprocessing.

The feature extraction phase is the most important phase in our model as the classification is based on features extracted during this phase. Features provide valuable information about different aspects or attributes of the wiki article, which are abstracted into a feature vector for the purpose of classification. The next section elaborates on the compact set of the proposed features for the outdatedness determination model.

The feature preprocessing phase refers to a process of representing the feature vector in a way that the machine learning classifier can process it.

In the last phase of our model, the generated feature vector is fed into a supervised machine learning algorithm in order to automatically classify wiki articles into their corresponding outdatedness classes. We adopt one of the best known supervised machine learning techniques, known as SVM, to evaluate the effectiveness of the feature vector we constructed [4, 9, 14].

5.1 Features

As previously discussed in the outdatedness determination model, a number of features are derived from various aspects of wiki articles by applying different feature extraction techniques. These features, as illustrated in Figure 2, fall into one of the following three categories: (i) article-based features, (ii) contributor authoritativeness features, and (iii) category-dependent features

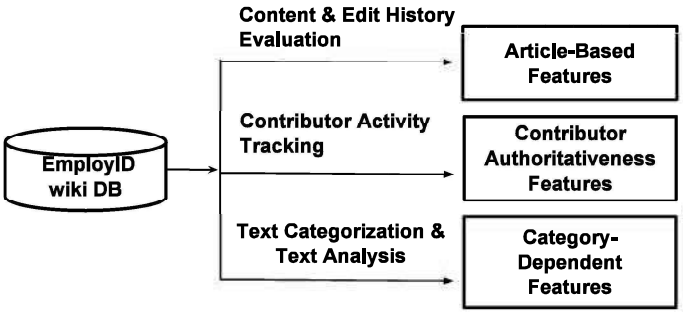

Fig. 2: Feature extraction workflow and categories

5.1.1 Article-Based Features

Article-based features, as depicted in Figure 2, are evaluated from a subset of two dimensions content and edit history.

Content features focus on the plain text of the article and are extracted from the textual content. In addition, content features ($f_1 - f_6$) also contain structural items of an article, as presented in Table 1. Dalip et al. [8] discover the capability of content features for predicting the quality of Wikipedia articles. Their work shows that applying content features, which are actually simple (low-level) indicators, can effectively improve the classifier performance when compared to other feature sets containing more complex features.

Edit history features represent the evolution of articles with respect to the timing of revisions, frequency of revisions, and community of editors. The features describing the timing of revisions are referred to as temporal features [5]. Edit history features are valuable features, which address different aspects of the *Knowledge Maturing Model* [15] such as stability, maturity, and collaboration. The edit history features ($f_7 - f_{14}$) are described as follows:

f_7: *(Age)* [2, 5, 16] This feature refers to the number of days since the articles creation time. At first glance, this feature can provide an indicator of outdatedness of the content of the article. On the other hand, it is observed that old articles are still valid as well.

Table 1: Content features

ID	Feature Name	Feature Description	Ref.
f_1	Character count	Total number of characters and ratio to the average character count of all articles within wiki	[2, 7]
f_2	Internal link count	Number of outgoing internal links (to other wiki articles)	[2, 16]
f_3	External link count	Number of outgoing external links	[2, 16]
f_4	In-link count	Number of incoming internal links (from other wiki articles)	[2]
f_5	Image count	Number of images	[2, 5]
f_6	Category count	Number of wiki categories assigned to an article	[2, 16]

f_8: *(Continuance of Edits)* [14, 17] This feature calculates the time passed between the creation of an article and the last revision, which shows the continuance of edit contributions.

f_9: *(Recent Changes Tracking)* The objective of this feature is to get an assessment of whether an article has been changed recently or not. This feature describes a numeric representation of performed revisions for the last four quarters.

f_{10} : *(Duration of Revisions to the First Revision)* [14] This feature refers to the time passed between the creation of an article and the creation of each revision. We consider the average duration as total duration divided by the number of revisions.

f_{11} : *(Edit Count)* [2, 16] This feature describes the total number of revisions.

f_{12} : *(Edit Rates)* [2] This feature calculates the ratio between continuance of edits f_8 and edit count f_{11}.

f_{13} : *(Author Count)* [2, 14, 17] This feature simply counts the number of distinct authors who contributed to the article.

f_{14} : *(Collaborative Feature)* The collaborative feature is a novel feature introduced in our work to investigate whether an article has been generated by relatively equal collaboration of authors or most generated by few of authors. We measure the percentage contribution of each author, then the standard deviation of these numbers is computed. This feature provides a valuable indicator regarding collaborative activities of authors.

5.1.2 Contributor Authoritativeness Features

The category of contributor authoritativeness features, as shown in Figure 2, evaluates the authority and activeness of wiki contributors. A number of studies [1, 11] have shown that articles with significant contributions from authoritative contributors are likely to be of high quality in terms of trustworthiness and reliability. The features ($f_{15} - f_{18}$) are represented in Table 2.

Table 2: Contributor authoritativeness features

ID	Feature Name	Feature Description	Ref.
f_{15}	Initiated pages count	Number of articles created by authors of an article	[14]
f_{16}	Participated pages count	Total number of contributions within wiki by authors	[2, 14]
f_{17}	Last contribution time	The last edit time of authors within wiki	—
f_{18}	Membership age	Time passed since registration of authors	[5]

5.1.3 Category-Dependent Features

As previously mentioned, articles in the EmployID wiki are correlated to categories. Categories are used to group related articles. In this section, we investigate novel feature sets to assess the outdatedness of articles within categories. In a different context, Weimer & Gurevych [20] suggest that predicting the quality of web forums within a specific topic can lead to a significant improvement in their classification performance. For instance, the category *Report* of the EmployID wiki, contains articles related to status reports of the project at different points in time. All articles in this category have the same structure such as an outline, review time, and probably an attached document. Considering articles within categories, we introduce the following features:

f_{19}: *(Category Indicator)* This feature indicates the category an article belongs to. We extract this feature by grouping related articles within categories. In the EmployID wiki, the most appropriate categories are assigned manually to wiki articles by authors. However, there are articles that are not assigned to a category (uncategorized articles). Therefore, we apply a text categorization task to group uncategorized articles referring to categorized articles into one predefined category.

f_{20}: *(Category Domain-Specific Feature)* This feature is defined as a numeric representation of domain-dependent entities in an article, considering the particular characteristics of the category it belongs to. This feature is calculated using text analysis techniques as part of the workflow shown in Figure 2.
A detailed description of all category domain-specific features would go beyond the scope of this paper. Preferably, we explain it for the *Report* category by way of example. Let $f_R = \{e_1, e_2, e_3\}$ be a feature vector with three entities assigned to a report article in the *Report* category, where

- e_1 defines whether the report was submitted due to an upcoming report or not.
- e_2 defines whether the outline was created for the report or not.
- e_3 defines whether the report notes are embedded in the wiki article as a document or not.

6 Experiment

In this section, we present the experimental setup and performed experiments to evaluate the performance of the outdatedness determination model.

6.1 Setup

The data of the EmployID Wiki is based on a real-life wiki in a project environment. A total number of 892 wiki articles are obtained, which were created within a time period of 2.5 years. First, a number of data preprocessing tasks are conducted before the dataset can be used effectively for classification. This includes removal of test pages, removal of all redirects referring to one particular page by multiple names or spelling, and the removal of wiki page records whose namespace is not zero. Wiki pages are grouped into collections called namespaces, which differentiate the purpose of pages at a higher level. Namespaces numbered zero contain real content, i.e., pages where main wiki articles reside. The final dataset after data preprocessing consists of 600 content articles.

In our experiment setup, two EmployID project members (as moderators) were asked to manually rate the outdatedness of articles, using scores between $1-5$. This range of scoring allows moderators to have more flexibility in judgments, as opposed to defining an absolute range of low, medium, and high classes. In order to measure whether the ratings are in agreement, Pearson correlation [13] is used, which is 0.625. This result shows that there is a positive association (agreement) between ratings of articles. The scores from the two moderators are averaged for the final score on each article. There are 9 possible average article rating scores, which are then distributed equally into three outdatedness classes so that $low = \{1, 1.5, 2\}$, $medium = \{2.5, 3, 3.5\}$, and $high = \{4, 4.5, 5\}$.

This manually rated dataset is used to train the classifier by learning important correlations between a set of features representing articles and their corresponding outdatedness classes. Table 3 shows statistics of the final dataset along with the number of rated articles in their respective outdatedness class.

Table 3: The EmployID wiki dataset

Total number of articles	600
Number of authors	50
Total number of users	79
Number of articles with low outdatedness	96 (16%)
Number of articles with medium outdatedness	126 (21%)
Number of articles with high outdatedness	378 (63%)

The SVM classifier is trained using LibSVM [6] to classify wiki articles into the three outdatedness classes. The SVM kernel is set to radial basis function

(RBF) with cost C and gamma γ parameters. A grid search is applied on C and γ parameters using cross-validation. The cross-validation is an iterative process, which can prevent the overfitting problem. The n-fold cross-validation evaluates the prediction accuracy of the model by partitioning the training set into n subsets of equal size. Sequentially, each subset is tested using the classifier trained on the remaining n-1 subsets. Thus, each instance of the whole training set is predicted once.

We improve the performance of the SVM classifier by performing data normalization. That is, all feature values are scaled to a range of $[0, 1]$.

6.2 Feature Selection

Sung & Mukkamala [18] show that the feature selection improves classification performance by searching for an optimal feature subset, which best classifies the training samples. In order to assess the outdatedness of articles within its categories, we need to identify the most important and relevant features, which efficiently construct the model describing articles in their related category. To this end, we employ a feature selection approach known as Sequential Forward Feature Selection (SFFS) [12]. This method selects the best single feature and then adds one feature at a time, which in combination with the selected features, minimizes the classification error.

6.3 Wiki Article Outdatedness Classification

We conduct three experiments to study the effect of using different feature sets on the classification performance of our outdatedness determination model. For all following experiments, the dataset is split into two disjoint subsets where 75% of the data is used for training, and 25% of the data is used as the test dataset.

In the first experiment, we employ a single SVM classifier, which is trained on all articles with different topics in the training dataset. For this experiment, we apply the feature set containing all features from article-based and contributor authoritativeness features ($f_1 - f_{18}$) as described in subsections 5.1.1 and 5.1.2.

In the second experiment, we investigate the effect of assessing the outdatedness of articles within categories by applying an alternate feature selection method, based on the category. Firstly, a text categorization task is applied to associate uncategorised articles into a fitting category. Secondly, for each category, the feature selection is used to select a subset of relevant features from article-based and contributor authoritativeness features ($f_1 - f_{18}$). Lastly, multi-SVM classifiers are employed in that the training of each classifier is done in parallel on articles of the same category using the selected feature set. The overall classification accuracy is computed as the average of the individual LibSVM classifier accuracy. All these steps are performed in an automated fashion. For example, once the category of an input article is determined, the related feature set is computed. Then, a classifier is invoked to evaluate the outdatedness class of that article.

The last experiment is set up as the previous experiment and in addition category-dependent features are added for articles in the same category. That is, for each category the selected feature subset from article-based and contributor authoritativeness features ($f_1 - f_{18}$) is combined with the category-dependent features ($f_{19} - f_{20}$). This results in category-dependent features and leads to increased performance of our model as shown in the following section.

7 Results and Discussion

In this section, we show and discuss the results of the three performed experiments for the classification of the EmployID wiki articles into the three outdatedness classes low, medium, and high. The performance for each experiment is presented in Table 4, which also shows the classifier performance in correctly determining the low, medium and high outdatedness classes.

Table 4: Summary of the classification performance of experiments

	Features	**Overall Accuracy**	**High**[2]	**Medium**[2]	**Low**[2]
1st experiment	$f_1 - f_{18}$	74%	92.70%	43.33%	37.5%
2nd experiment	$(f_1 - f_{18})^1$	85.33%	91.92%	70.37%	75%
3rd experiment	$(f_1 - f_{18})^1$ and $f_{19} - f_{20}$	92%	95.60%	87.88%	84.61%

[1]: Subset of features $f_1 - f_{18}$ is selected.
[2]: Accuracy for articles with high/medium/low outdatedness.

In the first experiment, we evaluate the accuracy of the proposed model with 18 features ($f_1 - f_{18}$), which are collected from article content, article edit history, and contributor authoritativeness. After training the LibSVM classifier on the normalized training set of 450 articles with the kernel parameters $C = 32$ and $\gamma = 0.5$ and 5-fold cross validation with the rate of 75.15%, the overall classification accuracy of 74% is obtained. That is, the classifier correctly classifies 111 articles from 150 articles of the test set. The results show that the classifier performs well in classifying articles with high outdatedness classes (92.70%, 89 out of 96), but relatively poor for articles with medium (43.33%, 13 out of 30) and low outdatedness classes (37.5%, 9 out of 24). After evaluating the misclassified articles, it is discovered that some articles, which were rated as low and medium outdatedness classes, are misclassified as highly outdated articles. In the following, we discuss reasons for this phenomenon. For example, the model may need to consider additional feature preprocessing tasks such as discretization, or to adopt another classifier such as decision trees, which can assist in improving classification performance. However, we believe that the misclassification problem, which occurs on the first experiment, is due to two main reasons. Firstly, as shown in Table 3, many articles fall into the class of highly outdated articles,

which can cause the classifier to overfit this class, leading to misclassification. Secondly, the proposed feature set ($f_1 - f_{18}$) can not effectively distinguish between outdatedness classes, because the feature set does not cover the category of articles.

In order to improve the performance of our model we conduct the second experiment, in which a feature selection method is applied on articles associated to the same category. Table 5 shows the most significant categories recognized in our experiment along with the subset of the selected features for each, which are identified by the sequential forward feature selection.

Table 5: Predominant categories in the EmployID wiki

Category name	**Articles in category**	**Selected feature subset**
Meeting	272	$f_1, f_3, f_7, f_8, f_9, f_{11}$
Report	60	$f_1, f_3, f_4, f_9, f_{10}, f_{12}, f_{13}, f_{14}$
Person	39	$f_1, f_2, f_4, f_{15} - f_{18}$
Partner	10	$f_2, f_4, f_{13}, f_{15} - f_{18}$
Tool	36	$f_1, f_3, f_4, f_7, f_9, f_{10}$
Task	10	f_7, f_8, f_9, f_{12}
Work packages	9	$f_1, f_7, f_9, f_{13}, f_{14}$
Other	164	$f_1, f_5, f_6, f_7, f_9, f_{10}, f_{12}, f_{13}, f_{14}$

In this experiment the overall accuracy of 85.33% is achieved. The results of the classifiers show that the average accuracy of an article being correctly classified as high outdatedness class is 91.92% (91 out of 99), as medium outdatedness class is 70.37% (19 out of 27) and as low outdatedness class is 75% (18 out of 24). This result indicates that assessing the outdatedness of articles within categories leads to significantly better classification performance, when compared to the first experiment. It is also observed that performing feature selection by removing irrelevant features of particular categories avoids the overfitting problem, which occurred in the first experiment. Moreover, these results are promising due to the reduction of the model complexity in terms of compute runtime for training and memory consumption because of smaller subsets of features.

In the following, we provide some concrete examples of correctly classified articles, in order to get a better understanding of how effective the selected feature subsets are for determining the outdatedness level of articles. These articles are selected from the *Meeting* category, which is the most popular category containing 272 articles. Table 6 shows six articles along with values of the selected feature subsets.

The articles represented in Table 6 are some instances of a monthly online meeting or annual review meeting of project members. These pages contain information about the meeting agenda and minutes, location information, links to materials, and lists of the participants. A discussion about the impact of

Table 6: Values of proposed features for the outdatedness classification of some articles from the *Meeting* category

	f_1	f_3	f_7	f_8	f_9	f_{11}	**Predicted Class**
1st Example	7175	8	287	73	[–,1,1,0]	10.04.16	Highly outdated
2nd Example	1784	2	311	59	[–,1,0,0]	03.03.16	Highly outdated
3rd Example	280	1	41	0	[–,–,–,1]	30.09.16	Medium outdated
4th Example	279	2	296	51	[–,1,0,0]	10.03.16	Medium outdated
5th Example	2210	0	105	7	[–,–,1,1]	04.08.16	Up-to-date
6th Example	341	2	41	34	[–,–,–,1]	03.11.16	Up-to-date

individual features is beyond the scope of this paper. Therefore, we only discuss some observations of manually analyzed features. After comparing the evaluated features of the first and the second examples with the features of the forth example, it is observed that these are old articles (f_7), which have not been changed at least in the last quarter (f_9). However, the first two examples are classified as highly outdated, and the forth one as medium outdated. We believe that there is a positive correlation between highly outdated articles and the character count feature (f_1). That means, an old article with long text, which has different sections, is more likely to be highly outdated because of the outdatedness of some of those sections. Another interesting observation is made here, when comparing the third example with the sixth one. Both articles are almost new (f_7, f_9), but their continuance of edits (f_8) differs significantly. While the third example has not been changed since the date of creation ($f_8 = 0$), the sixth example had some edits indicated by $f_8 = 34$. When comparing these two values, it is found that the sixth example can be considered more up-to-date, while the third example appears to be abandoned, and therefore is more likely to be outdated.

In the following, the results of the third experiment are discussed. In this experiment, we investigate the impact of category-dependent features ($f_{19} - f_{20}$) on the performance of the model, which provides the performance comparison between the results of this experiment and the previous experiment. The overall accuracy of 92% is obtained, which shows the best performance overall. As depicted in Table 4, this experiment performs best in classifying low and medium outdatedness classes when compared to the previous experiments. The results indicate that this sort of features can significantly contribute to the wiki article outdatedness determination.

8 Conclusions and Future Directions

In this paper, we propose a model and workflow to automatically determine the outdatedness level of knowledge in a wiki-based collaboration space. We make use of a supervised machine learning algorithm known as SVM to classify wiki articles into three outdatedness classes low, medium and high. Three different experiments are conducted to evaluate the impact of different feature sets on the

classification performance. After the initial evaluation of article-based, and contributor authoritativeness features, which resulted in a modest performance, it was discovered that the groups of articles in a same category have similar characteristics. This observation brings us to the assessment of outdatedness of articles within categories. To this end, the sequential forward feature selection method is applied on articles associated to the same category and multi-svm classifiers are trained in parallel. We introduce novel category-dependent features, which extract specific characteristics from groups of articles in categories. The overall classification accuracy of 92% is obtained, which shows the best performance overall.

Our results indicate that the proposed model and workflow can determine the outdatedness level of knowledge in wiki articles with significant accuracy. This provides an automated perspective on providing support for collaborative spaces. Based on that, we can think of presenting visual cues, advanced search ranking or similar support measures for users of collaborative spaces. An important insight is that this needs some limited domain knowledge on the category of articles, but then yields very good classification performance.

Our results suggest that certain features are good indicators for outdatedness, but collaboration practices that involve multiple different collaboration tools make it difficult to capture the phenomenon of outdatedness. The more functionality a single collaboration space can offer, the better the results will be. That means our proposed model and workflow is practically applicable to any collaboration space data that allows access to the document edit history, and user activities. Moreover, the documents of such collaboration spaces need to be rated ones manually for the training purpose.

Regarding future work, there are several axes can be suggested for further investiation. We plan to validate the results through user feedback. Moreover, we intend to apply the proposed model to other collaborative spaces. One interesting research direction would be to apply other supervised machine learning techniques, such as decision trees, neural networks, and Bayesian networks to classify the outdatedness of wiki articles and compare their performance. With regard to knowledge maturing, we intend to apply our model and workflow to the classification of relevancy in addition to the determination of outdatedness.

Acknowledgements

This work has been supported by the European Unions Seventh Framework Program for research, technological development and demonstration under grant agreement no. 619619.

References

1. Adler, B.T., Alfaro, L., Pye, I., Raman, V.: Measuring author contributions to the Wikipedia. In Proc. the 4th Int. WikiSym (2008)
2. Anderka, M., Stein, B., Lipka, N.: Predicting quality flaws in user-generated content: the case of Wikipedia. In Proc. SIGIR, 981–990 (2012)
3. Braun, S., Schmidt, A.: Wikis as a technology fostering knowledge maturing: what we can learn from wikipedia. In Proc. 7th Int. Conf. IKNOW 07, 321–329 (2007)
4. Chai, K., Hayati, P., Potdar, V., Wu, C., Talevski, A.: Assessing post usage for measuring the quality of forum posts. In Proc. the 4th DEST, 233–238 (2010)
5. Chai, K., Wu, C., Potdar, V., Hayati, P.: Automatically measuring the quality of user generated content in forums. In Porc. 24th Int. Conf. Adv. Artif. Intell., 51–60 (2011)
6. Chang, C., Lin, C.: LibSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm (2001)
7. Cozza, V., Petrocchi, M., Spognardi, A.: A matter of words: NLP for quality evaluation of Wikipedia medical articles. arXiv:1603.01987 [cs.IR] (2016)
8. Dalip, D., Goncalves, M., Cristo, M., Calado, P.: Automatic quality assessment of content created collaboratively by Web communities: a case study of Wikipedia. In Proc. JCDL09, 295–304 (2009)
9. Dang, Q-V., Ignat, C-L.: Measuring quality of collaboratively edited documents: the case of Wikipedia. In Proc. the 2nd IEEE CIC (2016)
10. Hu, M., Lim, E., Sun, A., Lauw, H., Vuong, B.: Measuring article quality in Wikipedia: models and evaluation. In Proc. CIKM07, 243–252 (2007)
11. Kane, Gerald C.: A Multimethod study of information quality in wiki collaboration. ACM Trans. Manage. Inf. Syst., vol. 2, no. 1, 4:1–4:16 (2011)
12. Kohavi, R., John, G.: Wrappers for feature subset selection. Artificial intelligence, 97, 273–324 (1997)
13. Lane, D. M.: Introduction to Statistics: An Interactive eBook. Chap.4, Values of the Pearson Correlation (2013)
14. Lee, J-T., Yang, M-C., Rim, H-C.: Discovering high-quality threaded discussions in online forums. J. Comput. Sci. Technol., 29:519–531 (2014)
15. Maier, R., Schmidt, A.: Explaining organizational knowledge creation with a knowledge maturing model. Knowledge Management Research & Practice, no. 1, 1–20 (2014)
16. Marzini, E., Spognardi, A.: Improved automatic maturity assessment of Wikipedia medical articles. In ODBASE Springer, 612–622 (2014)
17. Rizos, G., Papadopoulos, S., Kompatsiaris, Y.: Predicting News Popularity by Mining Online Discussions. SNOW/WWW, 737–742 (2016)
18. Sung, A.H., Mukkamala, S.: Identifying important features for intrusion detection using support vector machines and neural networks. In Proc. SAINT '03, 209–216 (2003)
19. Wanas, N., El-Saban, M., Ashour, H., Ammar, W.: Automatic scoring of online discussion posts. In Proc. the 2nd WICOW, 19–26 (2008)
20. Weimer, M., Gurevych, I.: Predicting the perceived quality of Web forum posts. In Proc. the Con. RANLP (2007)