Predicting Document Accesses with A Self-Enforcing Network

David Bergmann[1], Gregor Fuhs[2*], Christina Klüver[3]

[1] Institute for Computer Science and Business Information Systems (ICB), Universitaetsstr. 12, 45117 Essen, Germany.
[2] FIR e.V. at the RWTH Aachen University, Campus-Boulevard 55, 52074 Aachen, Germany.
[3] Institute for Computer Science and Business Information Systems (ICB), Universitaetsstr. 12, 45117 Essen, Germany.

* Corresponding author. Tel.: +49-241-47705-507; email: gregor.fuhs@rwth-aachen.de
Manuscript submitted February 5, 2019; accepted May 13, 2019.
doi: 10.17706/jcp.14.7.438-450

Abstract: Nowadays, construction projects are planned and carried out by large, interdisciplinary and mostly spatially separated teams. Project members have different data sets, which are often exchanged via complex and error-prone e-mail traffic. This leads to cumbersome and often chaotic data storage. In the following, a method with SEN as core element is described that is intended to improve data exchange through pre-planning data synchronization. The aim of the project is to develop a cloud-based collaboration software for the construction industry that sustainably increases the quality and speed of construction processes and reduces errors caused by old or incorrect data.

Key words: Self-enforcing network, document management, document access prediction.

1. Introduction

Large construction projects all over the world face the same problems. With interdisciplinary and often international teams, the complexity of communication and data exchange continues to increase [1]-[3]. On average, 35 different craft companies and up to 15 engineering and architecture companies are involved in a major project. Each team member involved in a project has their own data and systems to manage it [4]. In most cases no platform is used for data exchange, but the data exchange runs internally as well as externally via e-mails. The exchange of large amounts of data via e-mail leads to various problems that can delay the project. These range from limiting the file size of email attachments to incorrectly storing data. This also increases the risk of sending false or old documents. Modern software has revolutionized the areas of finance and accounting in companies through digitization. In contrast, technical processes and the exchange of data in large teams are only supported to a very limited extent by software. Especially the planning phase of construction projects and the associated document exchange is difficult to control due to high complexity. Large quantities of multiple documents stored in different versions and ineffective processing structures cause insufficiently documented, personal agreements and wrong decisions. This often leads to time and cost overruns.

The research project aims to develop software that eliminates these problems. The aim is to connect people, projects, data, places and languages in a very innovative but simple way. The development is divided into two parts: the technical part in which a 3-step hybrid cloud with proactive data

synchronization is developed and the implementation in which the system is further developed through user feedback (Fig. 1). This system will wrestle the problem of users waiting for files, which is not solved by Internet connections becoming faster since data volume increases accordingly.

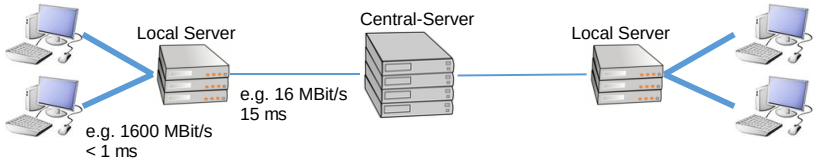


Fig. 1. 3-step hybrid cloud.

Current ECM and collaboration software for construction projects are a good first step to solve the described problems of the construction industry. However, due to a lack of standardization, cooperation between companies that often do not use the same collaboration software is still difficult. Project communication management systems and platforms offer a targeted solution but do not support process-oriented tasks or the connection to external applications. Therefore, the evaluation of the usability of these systems remains low and they cannot be used efficiently [5]. Cloud-based systems make teamwork more efficient [6]. Central data storage is the key to success. Precisely planned workflows and push notifications can also accelerate document exchange. Nevertheless, conventional cloud-based systems have the disadvantage of lengthy access processes, which slow down the processing of large amounts of data and cause errors. In addition, Internet speed varies greatly from region to region and mobile bandwidth in particular is a limiting factor for the use of known cloud based systems. None of the known products can really make a breakthrough in the efficiency of planning major projects in the construction industry.

2. Optimization of the File Transfer

The proposed research and development project aims to develop a new software to eliminate these problems. The goal is to link people, projects, data, places, languages and processes of construction projects to a highly innovative but simple way for the user. Today's ECM and collaboration systems for construction project management are a good first step to solve the described problems in construction companies [7]. However, due to a lack of standardization collaboration between other (not necessarily construction) companies with different systems is still a complex endeavor. Project-communication-management-software and platforms offer a more purposeful system for project communication but do not support process-oriented tasks or connections to external applications. Without the workflow-based data synchronization and external application connection, the usability rating of these systems is quiet low and therefore, these systems are not efficiently used and degenerate to a chaotic document storage [5], [8].

With cloud-based systems, collaboration in teams becomes more efficient [6]. Central data storage is a key point for efficient collaboration [9]. In addition, well-defined workflows and push-notifications can speed up the document processes. Nevertheless, conventional cloud systems have in many cases the disadvantage of lengthy access processes, which can slow down the processing of large amounts of data and generate errors. Furthermore, internet speed differ from region to region, especially mobile bandwidth on construction sites can be very low and therefore be a limiting factor. Therefore, large documents are often printed and sent via mail or by e-mail from one stakeholder to another. Sometimes, documents are only send to approve some minor changes or details.

None of the known products has made a real breakthrough in planning SMEs in the construction sector, in particular those with several branches. Instead of the promised time saving, the systems still have major

disadvantages:

- Manual, almost daily comparison of file versions of your own desktop files with the data on the server or in the cloud
- Manual, always repetitive long-term up- and download
- Group the files into their own structure
- No integration of specialized software such as CAD, BIM, FEM, etc.
- Manual work does not significantly reduce the incidence of errors

In this project, two methods are developed to increase the user's efficiency. On the one hand, the data transfer is accelerated by new methods of exchange (chapter 2); on the other hand, a neural network for prediction of document accesses is developed, whereby proactive document synchronizations are enabled (chapter 3).

In order to optimize data exchange, multiple methods were used. Initially, the system was converted to a stream transfer from the standard Windows file transfer. The transfer of many small files, which have to be confirmed individually, required a download time of 2.9 minutes on the simulated test track of 3687 KBit/s for 1000 files of 1 KByte each, i.e. they were relatively slow.

With the streaming method, on the other hand, the files are attached to each other, so that after the transfer the entire package only needs to be confirmed. In the case of an error-free transfer, this method is significantly faster, however, incorrect transfers require more time.

Since most document accesses are used for modification purposes of existing documents in the application case under consideration, it is advisable to modify the procedure here as well. Using the local delta method, the volume of files to be transferred can be significantly reduced. A delta is the difference between two files. The original file is buffered on the client so that it is available for later calculation. The user now edits this file. The delta of the modified files is then calculated with the original file so that only the delta has to be transferred to the server, which can generate the modified file from the original file and the delta (see Fig. 2).

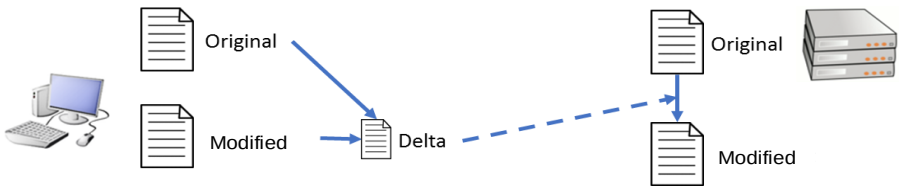


Fig. 2. Local delta method.

The reverse case is not so easy to implement, since the server would have to save the version of the document for each client in order to determine the appropriate delta. Therefore, in the Remote Delta Method, a "table of contents" with checksums of a file is determined, which the client sends to the server, from which the server calculates the delta (see Fig. 3). The deltas are compressed before they are transmitted.

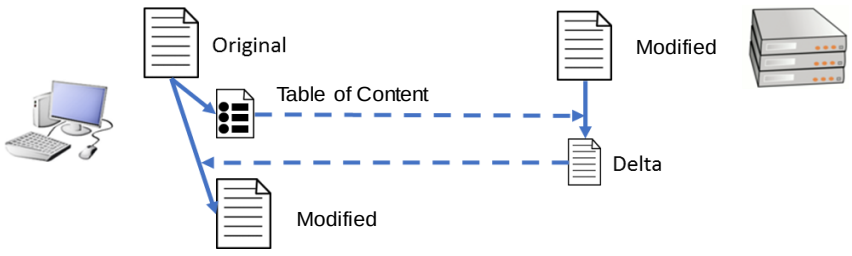


Fig. 3. Remote delta method.

Example scenarios:

- Initial download 90 MB AutoCAD file (at 3687 kbit/s)
  - Windows: 3.5 minutes
  - CCS Method: 12 seconds - of which are:
    - 3.8 seconds compression to 3.3 MB
    - 0.8 seconds extraction
    - (Remaining times: always mostly network times)
- Upload after small change
  - Windows: 8.5 minutes
  - CCS Method: 5 seconds - of which are:
    - 2.8 seconds Local delta calculation at 108 kb
    - (It was not compressed because 108 kb was below the current heuristic)
- 1000 files per 1 kb
  - Windows: Download 2.9 minutes - Upload 2.5 minutes
  - CCS Method:
    - Download: 1.5 minutes (incl. copy for local delta)
    - Upload: 1.1 minutes (for hard disk; faster for SSD)

3. Approach for a Proactive Data Synchronization System

The second approach to optimize the data transfer involves a proactive data exchange. Figure 4 visualizes the procedure of the system.

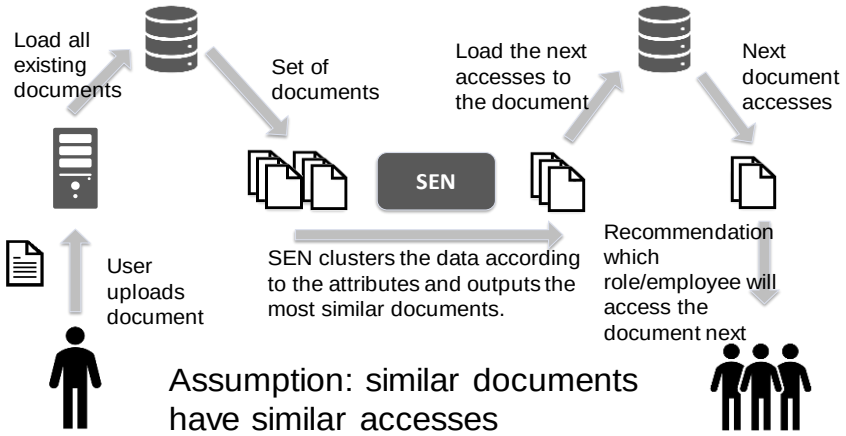


Fig. 4. Process of document prediction.

Each time a user uploads a document to the document management system (DMS), an event is created. This upload is an upload to a local instance of the server at a location. From there all existing documents are loaded into the cloud system. From the cloud system the prediction method clusters the data according to the attributes and outputs the most similar ones compared to the new uploaded one. All users who accessed the similar documents are likely to access the new document next and therefore the uploaded document should be transferred from the DMS to the local device of the recommended user. The central element for this cloud system is a self-enforcing network.

To train the prediction system an easy method is developed, shown in Fig. 5.

The training process starts each time a user accesses a file. This event triggers the system to predict a set of documents, which are most likely to be accessed next. When the user accesses a next file, the system gets a reward, if the file was part of the predicted set. Otherwise, the access information will be stored in a

database, which stores all false negative predictions. In regular cycles, the system uses the false negative events to optimize the parameters of the self-enforcing network and thus the prediction accuracy.

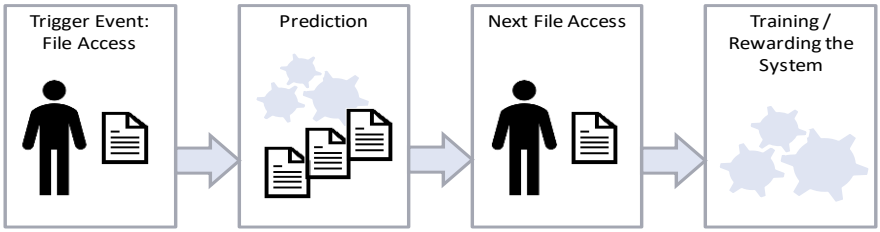


Fig. 5. Training method for the predicton system.

4. Self-enforcing Network

The following chapter briefly introduces the Self-Enforcing Network (SEN). The publications [10], [11] and [12] give a detailed insight into the functionality of SEN and contain application areas in which SEN has been successfully used. The SEN is a 2-layer neutral network introduced by Klüver and Klüver [13]. SEN calculates the similarity between already known objects and new objects. By determining the similarities, SEN can be used to categorize and cluster objects [12]. Each object can be represented by numerical attributes. The values thus indicate the degree to which an object has a certain characteristic. All already known objects are summarized in the so-called semantic matrix. The columns serve for the representation of the attributes and the rows of the matrix contain the individual objects. The attributes serve as input vectors, thus the neuronal net in the first layer has as many neurons as there are attributes. An output neuron exists for each object in the semantic matrix. Each neuron of the output layer is connected to each neuron of the input layer.

To create the net, the values of the semantic matrix are first recoded. Usually a unipolar or bipolar coding is chosen. Thus, the values after the coding are between 0 and 1 or -1 and 1. The coding is done as follows.

$$v_{norm} = \frac{v_{raw} - r_{min}}{r_{max} - r_{min}} * (n_{max} - n_{min}) + n_{min} \tag{1}$$

v_{norm} is the normalized value and v_{raw} is the raw value from the semantic matrix. The values r_{min} and r_{max} indicate the limits of the interval of the raw data and n_{min} and $n_{\max}$ indicate the limits of the interval to which the data is to be mapped. It is possible to weight the individual attributes differently using the Cue Validity Factor (cvf). To do this, the normalized values are multiplied by the cvf. If the cvf is less than one, the corresponding attribute does not have as much influence on the result. Conversely, if the cvf is greater than one, the corresponding attribute is weighted more strongly. The normalized data from the semantic matrix multiplied by the learning rate correspond to the weights between the input and output neurons. Thus, the weight w_{oa} of an attribute for an object results for the first iteration.

$$w_{ao} = c * v_{ao} \tag{2}$$

where v_{ao} is the normalized value of the semantic matrix for an object o and an attribute a. The value c is the learning rate specified by the user. If further learning steps should be necessary, the following applies for the calculation of the weights

$$w(t+1) = w(t) + \Delta w \tag{3}$$

with

$$\Delta w = c * w_{ao} \tag{4}$$

It has been shown that for many applications a learning rate $c = 0.1$ provides good results [14]. For the calculation of the values of the output neurons different activation functions can be used, which are shown in Table 1 [15].

Table 1. List of Activation Functions Available in SEN

Linear activation function	$a_j = \sum w_{ij} * a_i$
Linear mean function	$a_j = \sum \frac{w_{ij} * a_i}{k}$
Logarithmic linear activation function	$a_j = \sum \begin{cases} \lg_3(a_i + 1) * \mathrm{w_{i,j}}, & if\ a \geq 0 \\ \lg_3(\lvert a_i + 1 \rvert) * (-\mathrm{w_{i,j}}), & else \end{cases}$
Logistic activation Function	$a_j = \frac{1}{1+e^{-net_j}}$ with $net_j = \sum w_{ij} * a_i$
Enforcing activation function (EAF)	$a_j = \sum_{i=1}^{n} \frac{w_{ij} * a_i}{1 + \lvert w_{ij} * a_i \rvert}$

The generated network is used to determine the similarity between new objects and the objects in the semantic matrix. The attributes of the new object are used as input vectors. The calculated activations of the individual output neurons are used to determine the similarity. Since the output neurons represent the objects from the semantic matrix, a high activation of a neuron indicates the similarity between the corresponding element and the new object.

In addition to the activations, distances between the objects can also be calculated and used to determine the similarity. For this purpose not only the activations of the input vector are determined, but also the activations of the objects located in the semantic matrix. The distance of two objects is determined by the Euclidean distance of the activations.

5. Database and Model

The requirements to the model originate from the engineering office Schmidt. The procedure for the completion of the orders is characterized by project-based work. All documents that are created within the scope of the projects are stored on the own DMS, so that every employee has access to the documents. Access is provided by software installed on the employee's computer. The required document is downloaded from the DMS. After processing is completed, the updated document is reloaded onto the DMS. The document is locked on the DMS for the time of processing, so that processing by another user is not possible. Downloading with read-only permissions is nevertheless permitted. An offline availability would offer some advantages to the employees. Each time a document is used, the download results in short waiting times for the employee. Synchronization would eliminate these loading times. Employees would also be able to access the required data on the move. This is particularly helpful when viewing a building or when meeting customers, i.e. when a reliable internet connection is not guaranteed. In order to keep the costs for storage and the network load of offline synchronization low, it is necessary to develop a model that precisely predicts employee access to documents.

The log files recorded by the DMS between 04.12.2017 and 10.04.2018 serve as the basis for the evaluation of the developed model. All accesses made by the employees to the files are available. In addition, information about the documents is available. Information about the projects for which the documents

were created is also contained in the data record. Since SEN can only process numerical values, some data was converted at the beginning so that it can be processed. For example, the File Extension attribute is available for a document. To convert the file extensions to a numeric value, a value is assigned to each individual extension. Similar extensions have a similar value. For example, the .docx extension has a value of 52 and the .doc extension has a value of 50. The file extension .ppt, on the other hand, has the value 112. In addition, the extensions are categorized into classes. The extensions .doc and .docx are added to the Text document class and .ppt to the Presentation class. The classes in turn are also represented by an assigned numeric value. The roles of the employees are coded according to the same scheme. Cologne phonetics is used to encode the file paths. The method encodes words in numbers. Words with the same sound have a similar value.

All accesses are stored with a time stamp. For the evaluations, only accesses to documents created during the test period are considered. When a document is created, it is saved whether it was imported or created within the system. If a document does not show the information how it was created in the period, this is an indicator that it was created earlier and can therefore be filtered out.

Employee access to documents is considered implicit because access is an indicator of a user's interest in the document. The rating scale therefore has a binary value. A total of 68 active users access the DMS during this period. 27,451 documents are created which are accessed 51,368 times. Thus, each user accesses the system 755 times on average. On average, each document is accessed 1.9 times. However, a total of 16,776 documents are not accessed, except for the creation.

With only 1.9 accesses, and thus evaluations per document, the established techniques and methods of collaborative recommendation systems cannot be applied in this application context. It is not possible to make a statement about the similarities of users or documents on the basis of the evaluations. Too few overlaps exist to determine the similarity, i.e. users who have accessed the same document.

Looking at the context from a document-centric perspective, a static context arises because the document remains in the project for which it was created. In addition a specific goal can always be achieved by creating a document: an invoice as a request for payment, an engineering drawing as the basis for the construction object, or a project plan as milestone setting. The same steps are often taken to achieve the goal, which are guided through by the same roles. For example, on an engineering drawing an engineer, technical draughtsman or structural engineer is involved. This leads to the assumption that similar documents have similar accesses.

The algorithm for predicting accesses consists of three phases and is started by the occurrence of two events. Firstly, the algorithm is run through when a user creates a new document. The run generates initial recommendations for users who are interested in the newly created document. If a user accesses the document that was not recognized during the first recommendation, the algorithm is run through again. This is the second start event. If the algorithm has not recognized an access, it is assumed that the pattern was not recognized during the first run. Therefore, the recommendations are adapted with the second run. In the following, the document that triggers the algorithm is referred to as the current document. The first step is to load the data from the database. This does not load the documents themselves, but the information already described. In addition to the document and project data, information about the creation of the document is loaded. This includes the user who created the document and when it was created.

The input of the second phase consists of the document set created in the first phase. For this, the activations and the distances between all elements are calculated by means of a SEN. A hierarchical cluster analysis is used to cluster the documents. The algorithm used is similar to Single Linkage. For performance reasons, not all clusters are determined in the procedure used here, but only the relevant one, that is, the

one in which the current document is contained. The calculated distances or the ranks from the SEN serve as the basis for clustering.

The result of the clustering in phase two serves as the basis for phase three. The following accesses are loaded for all documents in the cluster and thus for all documents similar to the current element. The recommendation for which user the document is interesting consists of two branches. First, the current document is made available to the users who have direct access to the similar documents. On the other hand, the roles of the users are analyzed. The document is also made available to those users who have the corresponding roles in the project of the current document. Thus, the current context of the document is taken into account. From the domain view, this step makes sense because it is likely that a user who is also on the same project for which the document was created will also access it.

6. Results

The data described above were used to evaluate the model. A prototype was developed that implements the algorithm. For the evaluation, the regular operation of the DMS is simulated by iterating through the data in chronological order. At each iteration, it is checked whether the corresponding document was created or whether it is an access to an already existing document. If it is the first access, the algorithm is executed as described to create a recommendation for the document. If a subsequent access was recorded in the data, the system checks whether the model would have recognized this access.

The model contains a large number of parameters that can influence the result. Thus, the weighting of the available variables as well as the parameters of the SEN play a role. For the choice of a good parameter combination, the influence of the variables on the result is measured individually. Parameter optimization takes place on the basis of a three-week training period selected from the available data set. The parameter combinations that produce the best results are tested over the entire period to rule out overfitting and ensure that the results of the small period are transferable to a larger period. The process for optimization and evaluation is shown in Fig. 6.

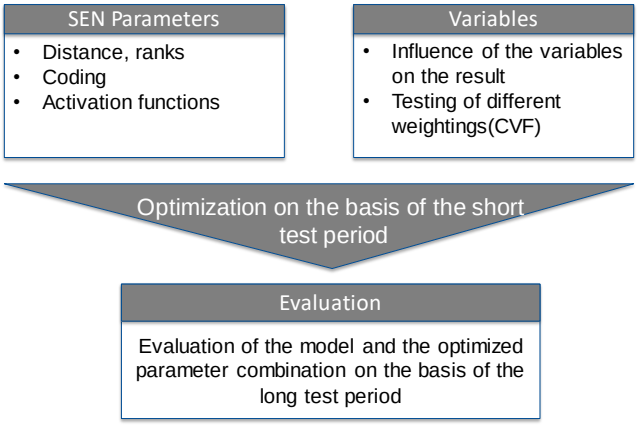


Fig. 6. Optimization and evaluation process.

SEN Parameters like distance, ranks, coding and activation functions as well as variables and combinations of them were varied to obtain the best possible prediction accuracy for the long test period. In the following, the parameters with the greatest influence are presented and the results of the simulation for the whole period are presented.

First, the influence of the different activation functions, encodings and the use of activations or distances on the result was examined. For each activation function, four simulations were performed based on the minor period. The four simulations differ in the different combinations of unipolar and bipolar coding and

use of the ranks and distances as similarity measure. When using the distances, it turned out that the coding [-1,1] yields better results than the unipolar coding. The activation functions have no great influence on the result. The precision varies by one percentage point and the recall varies by 2 percentage points.

On the other hand, the use of activations has shown that unipolar coding is more suitable than bipolar coding. There are no large deviations in the activation functions, the precision is 0.14 and the recall varies between 0.67 and 0.69. The exception is the logistic activation function. With this function, the recall is considerably higher at 0.85, but only achieves a precision of 0.04.

The learning rate and the learning steps have no influence on the result; therefore, the values 0.1 for the learning rate and 1 for the learning steps are used for further evaluation.

The influence of the variables on the result was also determined. The project number and the department leading the project proved to be good project-related variables. The document or access variables are the folder path, the number of access to the document, the date, the duration of access and the version of the document. The version can be divided into version phase, version release and version number. The weights of the three variables were given as 0.33, so that the version has a total weighting of 1. The other variables also have a weight of 1.

The results of the parameter optimization were used to create a simulation over the entire period. The results of the simulations are summarized in Table 2.

Table 2. Summary of the Results of the Simulations over the Entire Period

Activation	Similarity Measure	TP	FP	TN	FN	Precision	Recall	F-Value	MCC
Linear	Distance	15725	89206	1513117	5348	0.15	0.75	0.4167	0.3179
Linear	Activation	15700	116254	1486069	5373	0.12	0.75	0.3659	0.2785
Logistic	Activation	19126	383730	1218593	1947	0.05	0.91	0.205	0.1751

The table shows the best result of the runs, measured at the F-value and MCC, respectively for the use of the distance and activations for similarity determination. The linear activation function was used for both. The simulation using the distance as similarity measure achieves a recall of 0.15, so is more accurate than the simulation using the activations. With a constant recall, the result is also reflected by the F-value and the MCC. Simulations with a different activation function produce similar results. As already determined in the simulations using the smaller test period, the logistic activation function achieves an outlier. The parameter combination also achieves a very high recall with low precision in the simulation of the entire period. The low precision is reflected in both the F-value and the MCC.

Taking a closer look at the results, one can see that the precision and recall vary depending on the document type. Table 3 shows the result broken down by the most frequently used file extensions. The values come from the simulation where the linear activation function and the distance were used as a similarity measure.

At 57%, PDF documents are accessed most frequently. PDF files have a 1.4 percent better recall than the average. However, precision decreases by 3.5 percentage points. JPGs are the second most viewed files with 19%. A recall of only 59% is achieved here. The precision of 13% is also worse than the average. The model achieves an improvement in both the recall and the precision of the DOC or DOCX, XLS or XLSX documents and DWG drawings. All file extensions achieve a recall of over 80% and in some cases an precision of 40%.

Table 3. Representation of the Recall and Precision of the File Extensions

File Extension	True Positive	False Negative	Recall	Precision
pdf	9166	2827	76,43	11,49
jpg	2335	1627	58,93	13
doc	1533	191	**88,92**	40,98
dwg	629	152	80,54	20,42
docx	646	113	85,11	29,38
xlsx	518	88	85,48	40,22
xls	189	35	84,38	35,06
zip	146	49	74,87	12,09
msg	182	12	**93,81**	51,41
tif	109	56	66,06	16,1
ppt	8	60	11,76	4,44
pptx	22	16	57,89	13,92

7. Discussion of the Results

The results show that for some file types the accesses can be predicted well with the developed model. The results reflect initial expectations. Office files or engineering drawings often have a specific goal to achieve by creating the file. The same people often work together to achieve the goal.

The assumption made at the beginning can be confirmed by the present results, since the model achieves a high recall as well as precision for the file types mentioned. The model to create the forecast thus uses the patterns in the data. PDF files, on the other hand, are created from the aforementioned documents. This completes the processing. Anyone who is interested in the content of the PDF could access the document, making the target group for a PDF larger. This could explain the worse result compared to Office files. One reason for the low recall of JPGs may be their diversity. Construction sites are often photographed in order to create images for documentation purposes. A large number of photos are created, some of which are no longer used at all after uploading. Others are used for the creation of reports or shown to customers as an example. Which photo is used for a report cannot be predicted with the given data, because the metadata hardly differ from each other. Photos are often stored in a folder called Pictures or Photos. The consecutive number created by the camera is used to name the files.

The model can be used to apply offline synchronization only to relevant documents. This would offer the user the advantage of no longer having to download the required documents, thus eliminating waiting times. Irrelevant data would not be downloaded. This saves the memory on the local devices and reduces the network load compared to a complete offline synchronization. However, in real-world use, further questions arise that go beyond the creation of the recommendations. In order to ensure that the current versions of the files are available on every device, all local copies must be updated after a document has been edited. It is also necessary to determine when a document is no longer needed and can be deleted from the local device, as the storage requirement would continue to grow with permanent local storage.

The model provides good results for the available data. In order to refine and improve the model, future work should focus on two areas. On the one hand, the extension of the variables is necessary to achieve results that are more precise. So far, only the variables stored in the DMS have been used. However, it may be useful to analyze the content of the documents. So the content in text documents could be informative or

an image analysis to improve the results of the JPG. On the other hand, the data that are loaded into the SEN should already be pre-sorted, since the database continues to grow with the use of the system. Thus, the creation of recommendations becomes more time-consuming, since all past data is used for the preparation of a recommendation. Here upstream cluster procedures could limit the input data in order to guarantee the performance also with intensive use of the system.

8. Advantages of Proactive Document Distribution

The advantages of offline storage are offset by costs. These costs result from downloading incorrectly recommended documents. It does not make sense to evaluate the advantages and disadvantages based on the number. The number does not provide any information about the memory required or the duration of the download. Table 4 shows the ten employees with the most accesses, measured by data volume. The second column of the table shows the cumulative data volume of the documents that the employee accessed in the period. The third column displays the data volume of the recognized accesses. This data can then be downloaded in advance and made available to the employee. This means that 4.1 GB of the 5.1 GB of data required for employee 1 can already be downloaded in advance. This means that only one gigabyte of data needs to be reloaded. The file sizes stored in the database are used to calculate the data volumes. These are also used for the calculation of the required memory for the provision of all recommendations. However, the used system has the property to save the file size only with the second access. This means that after creating a document, the size of 0 bytes is specified for the document. This means that the database only contains information about the size of a document to which a subsequent access also exists. The size cannot be determined for the documents that were created and have no further call.

The fourth column shows the memory requirements of the recommendations. The calculation is based on the values from the database. Thus the sizes of the documents are missing, which were not accessed further, but were determined as helpful for the user. The memory requirement of these documents is determined in the fourth column. The calculation is based on the average sizes of the different file types.

Table 4. Evaluation of the Detected Amount of Data

User	Size of all accesses (in MB)	Size of predicted accesses (in MB)	Size of recommendations (in MB)	Size of missing recommendations (in MB)
1	5165	4128	4884	6594,3
2	3558	2482	3510	1080,64
3	3089	2326	3331	7283,69
4	2534	1509	5315	6560,81
5	2477	2082	3225	5433,94
6	2378	1580	2965	6912,56
7	2359	1944	1099	5548,96
8	2133	1607	4991	4449,08
9	1840	1600	4329	4743,03
10	1755	1580	2023	1849,75

9. Conclusion

The proposed approach of system for proactive data exchange would support the collaboration on complex construction projects and help the building industry to climb a new level regarding digitalization. Challenging projects would be processed with considerably fewer errors and therefore higher accuracy regarding deadlines. However, the implementation and acceptance within the industry is still a challenge.

To unleash the full potential, software systems need to get a higher acceptance among all stakeholders within the building industry. The current results show that SEN is a good approach to prediction. The first tests showed very good results for prediction and data load, which have to be confirmed in the next steps with further data to avoid an overfitting of the parameters. However, it can be assumed that the problem formulated can be solved using SEN as a prediction tool.

References

[1] Rueppel, U. (2007). *Vernetzt-Kooperative Planungsprozesse im Konstruktiven Ingenieurbau*. Berlin: Springer.

[2] Wu-Lee, C., & Hwang, G. H. (2013). *Workflow Definition by Cloud Collaboration*. *Proceedings of the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing.*

[3] Fraunholz, B., Unnithan, C. R., & Chamberlain, J. (2003). Project communication management in Australia, Germany and India — A cross cultural study. *Proceedings of International Conference on Information Technology and Organizations: Trends, Issues, Challenges and Solutions. Conference: Information Resources Management Association* (2002: Philadelphia, Pa.).

[4] Guenthner, W. A., & Borrmann, A. (2011). *Digitale Baustelle — Innovativer Planen, Effizienter Ausführen*. Heidelberg: Springer.

[5] Klauer, T. (2005). *Eine Prozessorientierte Kooperationsplattform für Bauprojekte auf Basis Eines Internetbasierten Workflow-Managements*. Shaker.

[6] Lin, C., Wayne, Y. W. C., & Wang, J. (2014). *Cloud Collaboration: Cloud-Based Instruction for Business Writing Class*. World Journal of Education, Sciedu Press.

[7] Ma, K., Dawood, N., & Kassem, M. (2017). *BIM for Manufacturing: A Case Study Demonstrating Benefits and Workflows and An Approach for Enterprise Application Integration (EAI)*. *Proceedings of the 16th International Conference on Construction Applications of Virtual Reality*. Hong Kong.

[8] Gessinger, S., & Bergmann, R. (2015). Flexible process-aware information systems deficiency management in construction. *LWA.*

[9] Kraus, R., Fillinger, S., Tolksdorf, G., Minh, D. H., Merchan-Restrepo, V. A., & Wozny, G. (2014). Improving model and data integration using mosaic as central data management platform. *Chemie Ingenieur Technik, 86(7).*

[10] Klüver, C. (2012). Solving problems of project management with a self enforcing network (SEN). *Comput Math Organ Theory, 18(2).*

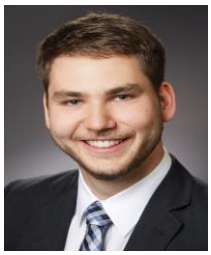
[11] Klüver, C. (2016). Self-Enforcing Networks (SEN) for the development of (medical) diagnosis systems. *Proceedings of 2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 24-29). Vancouver, Canada. Piscataway, NJ: IEEE.

[12] Klüver, C. (2016). Steering clustering of medical data in a Self-Enforcing Network (SEN) with a cue validity factor. *Proceedings of IEEE Computational Intelligence Society (Ed.): 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*.

[13] Klüver, C., & Klüver, J. (2013). Self-organized learning by self-enforcing networks. *Proceedings of International Work-Conference on Artificial Neural Networks*.

[14] Klüver, C., Klüver, J., & Schmidt, J. (2012). *Modellierung Komplexer Prozesse Durch Naturanaloge Verfahren. Soft Computing und Verwandte Techniken*. Wiesbaden: Springer Fachmedien Wiesbaden.

[15] Klüver, C., Klüver, J., & Zinkhan, D. (2017). A self-enforcing neural network as decision support system for air traffic control based on probabilistic weather forecasts. *Proceedings of 2017 International Joint Conference on Neural Networks (IJCNN), IEEE*.



David Bergmann is a master student of business informatics at the University of Duisburg-Essen. He completed his bachelor's degree at the Westfälische Wilhelms Universität Münster and obtained his B.Sc in 2015. During his studies, he focused on collaboration environments using cloud based solutions. David Bergmann has also been intensively involved in the field of neural networks since his master's degree. The focus is on the development of models using the Self-enforcing network.

Gregor Fuhs studied mathmatics at the RWTH Aachen University, where he obtained his master degree in 2015, focusing on numerics and informatics. Subsequently, he dedicated his doctoral thesis to the propagation of errors in information flows. In particular, he studied methods for information and data analysis. Additionally, he researches the subject of enterprise content management and advises companies on the implementation of an ECM concept and other information logistics issues.

Christina Klüver (Dr. phil., habil.) studied educational science and computer science. She obtained the Ph.D. in communication science and the venia legendi in computer science. She is lecturer in computer science at the University of Duisburg-Essen, Germany, in the Faculty of Economics and Business Administration. As a member of the research group COBASC (Computer Based Analysis of Social Complexity) her research interests are in applying models of nature analogous programming techniques like neural nets, cellular automata, and evolutionary algorithms to technical, social, and economical problems.