Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2007 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2007

A Method for File Valuation in Information Lifecycle Management

Lars Turczyk

Marcel Groepl
TU Darmstadt

Nicolas Liebau

Ralf Steinmetz
Technical Universtity Darmstadt

A Method for File Valuation in Information Lifecycle Management

Lars Arne Turczyk
Technische Universität Darmstadt
KOM – Multimedia Communications Lab
Merckstr. 25
D-64283 Darmstadt
Germany
lars.turczyk@siemens.com

Marcel André Gröpl
Technische Universität Darmstadt
KOM – Multimedia Communications Lab
Merckstr. 25
D-64283 Darmstadt
Germany
marcel.groepl@kom.tu-darmstadt.de

Nicolas Liebau
Technische Universität Darmstadt
KOM – Multimedia Communications Lab
Merckstr. 25
D-64283 Darmstadt
Germany
nicolas.liebau@kom.tu-darmstadt.de

Ralf Steinmetz
Technische Universität Darmstadt
KOM – Multimedia Communications Lab
Merckstr. 25
D-64283 Darmstadt
Germany
ralf.steinmetz@kom.tu-darmstadt.de

ABSTRACT

ILM is based on the idea that in an enterprise different information have different values. Valuable information is stored on systems with a high quality of service (QoS). The value changes over time and therefore migration of information is required to cheaper storage systems with a lower QoS. Automated migration makes ILM dynamic. Such automation requires storage systems to understand what files are important at what time so that right policies can be applied. In this point ILM nowadays lacks information valuation methods.

This paper looks at how the value of a file can be measured. Different from traditional methods using metadata leading to a classical decimal-value we show how the value can be derived using a probabilistic method. Here the value of a file is calculated from usage information and expressed as a "probability of further use". This is a new method which allows valuation depending on the future importance of a file.

Feasibility of the new method is verified by generating file migration rules for ILM.

Keywords

Information Lifecycle Management, File Valuation, Access Behaviour, File Migration Rules.

INTRODUCTION

Information Lifecycle Management (ILM) stores files according to their value. Therefore file valuation is a very important task in the ILM environment. The question is "How is the value of a file measured?".

Storage Network Industry Association (SNIA) proposes measuring the value as an amount of money (Peterson, 2004). This method is quite intuitive, but not feasible for environments with a large number of files. Other methods express the value as a decimal-value (Chen, 2005). These methods need metadata to calculate the value. The problem, though, is collecting and updating the metadata over the lifecycle of the file, which generates great effort. Other methods express the value as the period of not being accessed (Tanaka, Ueda, Aizono, Ushijima, Naitoh and Komoda, 2005). These methods are too static and do not reflect the value changes over time adequately.

A feasible, dynamic method for file valuation is needed without considering metadata. Therefore we demonstrate how the value can be derived using a probabilistic method. Here the value of a file is calculated from usage information and expressed as a probability of further use. It expresses the future importance of a file, thus making it easy to decide on storage location.

To create the method we conducted a case study at a German blue-chip company, where the access behaviour of various Microsoft (MS) Office files was observed. The paper presents the case study and the statistical methods used to determine the

probability of further access. File migration rules for ILM can be generated using the results of the case study. The paper ends with the definite calculation of access probabilities and the application of migration rules.

The essence of this paper is as follows:

1. We present a case study and derive distribution functions to describe the access behaviour of different MS Office file types.
2. We use the access behaviour to valuate files.
3. We show that this probabilistic method of file valuation is feasible for ILM and we apply the method for generating file migration rules.

RELATED WORK

Migration of files from more expensive storage to less expensive storage has been studied as far back as early1980s. These studies (Smith, 1982 and Lawrie, Randal, and Barton, 1982) concluded that a file selection algorithm based on file age and size results in a minimum amount of file recall occurrences and in optimal storage utilization.

Long-term access behaviour was examined already in 1992 by Strange (1992), 1998 and 1999 by Gibson et al. (1998 and 1999), and 2004 by Schmitz (2004). The observed time periods varied from 84 days by Strange and Schmitz to 280 days by Gibson et al. The data examined by the authors originated from UNIX file systems at German (Schmitz 2004) and American (Gibson and Miller 1999, Gibson et al. 1998, Strange 1992) research centres. The data used for this paper originate from a Microsoft file system at a German blue-chip company. Moreover, the evaluated time periods are even longer: The complete life cycle of files aged up to 1771 days (more than 4 1/2 years) are analysed below. Strange derived the ”least-recently used algorithm”. This shifts those files that have not been used for the longest time first. Gibson and Miller examined the so-called ”file-aging algorithm”, which also considers a previously calculated migration value besides the file size and relapsed time since the last use (Gibson and Miller 1999, Gibson et al. 1998). This migration value should measure the intensity of use with respect to time: It increases on a day, on which the corresponding data was used, and reduces for each day of non-use.

The described publications do not determine statistical distribution models, whereby they differ fundamentally from this paper. Usage information is used for valuation in other system domains as well. Google uses PageRank algorithm to rank the importance of a web page (Page, Brin, Motwani and Winograd, 1999, Ridings and Shishigin, 2002). A page is ranked based mainly on how many other pages are linked to it. Such links represent a form of usage. They indicate how many other pages are using that particular page. Caching algorithms often rely on data usage information to determine what data are important and hence what to cache in buffers in file systems, databases, and storage controllers (Denning, 1968 and 1980, Effelsberg, and Haerder, 1984). These algorithms cannot be directly applied to our problem due to different design purposes and different target data.

Today ILM is a strict focus of research. The main results are found in the field of “how” ILM works, i.e. most research was done in the field of procedures and policies. Vendors presented their understanding of ILM (Reiner, Press, Lenaghan, Barta and Urmston, 2004). Turczyk, Berbner, Heckmann and Steinmetz (2006) gave a formal definition usable for ILM abstraction. Beigi, Devarakonda, Jain, Kaplan, Pease, Rubas, Sharma and Verma (2005) and Tanaka et.al. (2005) offered proposals for policy description of ILM. Last but not least, Chen (2005) focused on the valuation of files. His approach differs from ours, because does not use probabilities for valuation.

ASCERTAINMENT AND DESCRIPTION OF DATA

The examined database contains approx. 150,000 files and their access protocols. For the following statistical analysis a random sample of 1000 files was extracted. The following data is known for each file: File type, file size, date and time (accurate to one minute) of file creation and the date, time and type of the individual accesses. This information was conditioned after the random sample extraction using MS Excel® to make the data needed for each analysis available. The programs R and MATLAB® were used for the data analysis.

Description of the random sample attributes

Tables 1 to 6 characterize the random sample by illustrating the frequency distributions of the number of accesses per file, the size of the files, the size of the accesses, the age of the files, as well as the file types and access methods.

Number of accesses	[1;2)	[2;3)	[3;4)	[4;5)	[5;10)	[10;20)	[20;50)	[50;100)	[100;200)	[200;292)
Number of files	307	152	99	79	209	77	53	14	6	4

Table 1. Number of accesses per file

Size of files	[1kB;10kB)	[10kB;50kB)	[50kB;100kB)	[100kB;500kB)	[500kB;1MB)
Number of files	22	265	158	267	108
Size of files	[1MB;2MB)	[2MB;5MB)	[5MB;10MB)	[10MB;50MB)	[50MB;115MB)
Number of files	81	48	36	12	3

Table 2. Size of the files

Size of accesses	[1kB;10kB)	[10kB;50kB)	[50kB;100kB)	[100kB;500kB)	[500kB;1MB)
Number of accesses	169	2357	1228	1408	1458
Size of accesses	[1MB;2MB)	[2MB;5MB)	[5MB;10MB)	[10MB;50MB)	[50MB;115MB)
Number of accesses	440	426	322	65	38

Table 3. Size of the accesses (size of the accessed file in each case)

Age of files	[0;1 w)	[1 w;1 m)	[1 m;¼ y)	[¼ y;½ y)	[½ y;1 y)
Number of files	7	37	87	109	231
Age of files	[1 y;1½ y)	[1½ y;2 y)	[2 y;3 y)	[3 y;4 y)	[4 y;5 y)
Number of files	247	80	138	36	28

Table 4. Age of the files (w = week, m = month, y = year)

File type	doc	xls	ppt	Pdf	zip	msg	miscellaneous
Number of files	335	185	164	140	41	24	111

Table 5. File types

Access type	Version Fetched	View	Version Added	Move
Number of accesses	3657	1519	1392	438
Access type	Reserve	Unreserve	Permission Changed	Miscellaneous
Number of accesses	256	247	200	202

Table 6. Access types

The 1000 files in the sample were accessed a total of 7911 times between their respective creation and their extraction for the random sample (see Table 1). Care must be taken when considering the number of accesses that the first access to a file in the

examined database is logged at the time of its creation. As a result, 307 files were not accessed one single time after their creation. After discounting these "unused" files, most of the files, i.e. 152, were accessed only once after the creation date.

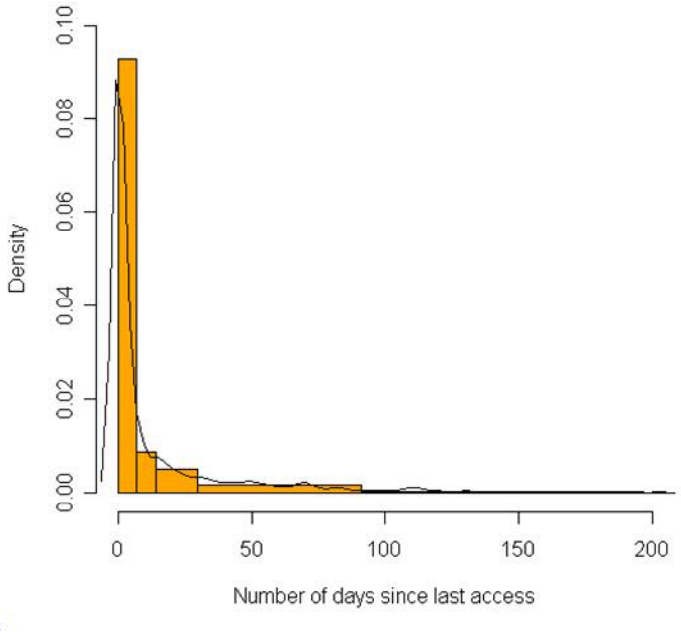
The file types doc, xls, ppt, pdf and zip are contained most frequently in the sample (see Table 5). The file types avi, cfg, csv, cti, dot, exe, gif, htm, jpg, log, mdb, mmap, mmp, mp3, mpg, mpp, pps, pst, rtf, sql, tif, trc, txt, vsd, vss, wav, wbk, wf2 and xml fall into the category "miscellaneous".

Most accesses to files in the sample, i.e. 46.22 % of 7911, are of the "version fetched" type (see Table 6). The access types "View" and "Version added" are represented with 19.20 % and 17.60 % at second and third place. Other frequently occurring access types are "Move", "Reserve", "Unreserve" and "Permission changed". The noticeably more seldom access types under "Miscellaneous" are "Attributes Changed", "Rename", "Copy", "Version Deleted", "Alias Created" and "Generation Created".

Expired time since last access

The expired time since the last access are examined in this section. All of the first accesses to the 1000 files drop out of this consideration, whereby the number of accesses is reduced to 6911. 307 files that experienced only this one access drop out fully, so that only 693 files are analysed from here onwards.

The histogram and density curve in Figure 1 have a very pronounced steep rise to the left: Almost half of all accesses occurred within one day of the last access. The averaged expired rime since the last access is 33.71 days.



Time since last access	[0;1 w]	(1 w;2 w]	(2 w;1 m]	(1 m;¼ y]	(¼ y;½ y]
Number of accesses	4485	419	568	785	301
Time since last access	(½ y;¾ y]	(¾ y;1 y]	(1 y;2 y]	(2 y;3 y]	(3 y;3,7 y]
Number of accesses	131	125	66	24	7

Figure 1. Expired time since last access

The density curve approaches a very low density within only half a year: Only 5.11 % of the accesses occurred after expiry of a half year after the last access and after three quarters of a year only 3.21 %. Expired times since the last access can thus be quoted, e.g. three quarters of a year, after which the probability of further accesses is very low. The expired time since the last

access can therefore be used as the characteristic for a migration rule and will therefore be analysed with advanced statistical methods.

Correlation analysis

The relationship between the attribute "Number of days since the last access" and the attributes "File size" and "File age" had been examined. The result is that the considered relationships are not suitable for use in a migration rule, since only a weak co-relationship exists and a specification using a regression analysis is impossible.

THEORETICAL DISTRIBUTION MODELS

The definition of the random variable "Number of days since the last access" is as follows: Let (Ω, *F*, *P*) be an arbitrary probability space. The operator $X:\Omega \to \Re$ is a random variable, when

$$A:\left\{\omega \in \Omega \middle| X(\omega) \le x\right\} \in F \quad \forall x \in \Re$$

where

F: σ-algebra

P: Probability

Ω: Population, here: Set of all accesses

ω: Arbitrary element from Ω, here: Access

X(ω): Here: Number of days since the last access

Only "life span distributions" are to be considered to model the random variable X. According to Hartung (2005) and Schlittgen (2003), these include

- The exponential distribution,
- The Weibull distribution,
- The Raleigh distribution,
- The Gamma distribution,
- The Erlang distribution and
- The IDB distribution (Hjorth distribution).

One establishes after a series of Q-Q-Plot tests that only the Weibull and the Gamma distribution are suitable.

The χ^2 adaptation test sheds light on whether the random variable X can be viewed as being Weibull-distributed or gamma-distributed. The hypotheses to test the assumption of a truncated Weibull distribution are:

H_0: The random variable X is descended from a truncated We(0.30;6.67) distribution

against the alternative

H_1: The random variable X does not emanate from a truncated We(0.30;6.67) distribution.

It is tested to significance level 0.001.[1] The test term T is 307.89 and the following applies:

$$T > \chi^2_{47;0.001} \Leftrightarrow 264.07 > 82.72$$

This means, H_0 must be rejected.

To test the assumption of a truncated gamma distribution, the hypothesis

H_0: The random variable X is descended from a truncated Ga(0.14;268) distribution

[1] The significance level α is the probability, with which a correct zero hypothesis is rejected. The closer α approaches zero, the more likely one retains the zero hypothesis. (Sachs, 2004)

against the alternative hypothesis

H_1: The random variable X does not emanate from a truncated Ga(0.14;268) distribution

is tested to significance level 0.001. One obtains a value of 203.73 for the test term T. The rejection range is as follows:

$$T > \chi^2_{47;0.001} \Leftrightarrow 203.73 > 82.72$$

This means, H_0 must be rejected.

On the basis of the test values in the case of the Weibull distribution and in the case of the gamma distribution, one can declare that the adaptation to the empirical distribution is slightly more successful with the latter, but that the zero hypothesis must nevertheless be rejected. The application of more advanced statistical methods is thus necessary.

CONSTRUCTION OF A MIXED DISTRIBUTION FUNCTION

The mixture of a discrete and a continuous distribution function will be laid down and then tested in this chapter.

The probability p, that the waiting time W or the number of days since the last access equals zero is calculated to $p = \frac{1079}{6911}$.

$F_{W>0}(x)$ is the distribution function of the random variable X: "Number of days since the last access", however under the condition that W > 0. In other words: $F_{W>0}(x)$ is the distribution function of the 6911-1079=5832 observations of more than zero days since the last access. Therefore the following mixed distribution function applies:

$$F(x) = \frac{1079}{6911} \cdot \mathbf{1}_{[0,1327)}(x) + \frac{5832}{6911} F_{W>0}(x) \tag{1}$$

Mixed distribution function with truncated Weibull distribution

The mixture of a discrete and a continuous distribution function will be tested in this section, whereby a truncated Weibull distribution is assumed for the continuous distribution function. The following terms apply:

- FW>0(x): Distribution function of the waiting times greater than zero (here: Weibull distribution function)
- F*W>0(x): Truncated variant of FW>0(x)
- F(x): Distribution function of the total waiting

The Weibull distribution is truncated to the value range $[6.94 \cdot 10^{-4}; 1327]$, because the minimum number of days since the last access is $6.94 \cdot 10^{-4}$ days (one minute) and the maximum number is 1327 days. The mixed distribution function is now as follows:

$$F(x) = \frac{1079}{6911} \cdot \mathbf{1}_{[0,1327)}(x) + \frac{5832}{6911} F^{\star}_{W>0}(x) \tag{2}$$

where

$$F^{\star}_{W>0}(x) = \frac{F_{W>0}(x) - F_{W>0}(6,94 \cdot 10^{-4})}{F_{W>0}(1327) - F_{W>0}(6,94 \cdot 10^{-4})} \qquad 6,94 \cdot 10^{-4} \le x \le 1327$$

The optimised parameters of the truncated Weibull distribution are $\hat{\alpha} = 0.33$ and $\hat{\beta} = 9.90$. The χ^2 adaptation test is now performed. The hypothesis

H_0: The random variable X is descended from a population with the mixed distribution according to equation 2 with the truncated Weibull distribution function $F^*_W(x)$

is tested to significance level 0.001. One obtains a value of 236.29 for the test term T. The rejection range is:

$$T > \chi^2_{48;0.001} \Leftrightarrow 236.29 > 84.04$$

This means, H_0 must be rejected. The test value could be reduced slightly in comparison to the truncated Weibull distribution, i.e. from 264.07 to 236.29, but in spite of this, a positive test result also could not be achieved in the case of the mixed distribution function consisting of a step function and truncated Weibull distribution.

Mixed distribution function with truncated gamma distribution

Analogue to the previous section, a mixed distribution function will be tested here, however in this case, a truncated gamma distribution is assumed for the continuous distribution function.

The gamma distribution is also truncated to the value range $[6.94 \cdot 10^{-4}; 1327]$. The mixed distribution function to be tested corresponds to equation 2 with the term $F_{W>0}(x)$ for the non-truncated gamma distribution and $F^*_{W>0}(x)$ for the truncated gamma distribution.

The optimised parameters of the truncated gamma distribution are $\hat{\alpha} = 0.14$ and $\hat{\beta} = 260.0$. The zero hypothesis of the χ^2 adaptation test is again:

> H_0: The random variable X is descended from a main unit with the mixed distribution according to equation 2 with the truncated gamma distribution function $F^*_{W>0}(x)$.

It is tested to significance level 0.001. One obtains a value of 200.58 for the test term T. The rejection range is:

$$T > \chi^2_{45;0.001} \Leftrightarrow 200.58 > 80.08$$

H_0 must therefore be rejected.

The result of this section is that a positive test result was not achieved with a mixed distribution function consisting of a step function and a truncated Weibull or gamma distribution. According to this, it is not possible to specify a distribution function for the whole sample and therefore distribution assumptions for partial samples will be examined.

Division of the sample according to file types

The sample divided according to file types will be examined below according to the approach described in the previous sections. The partial sets are specifically 1358 observations of accesses to doc files, 2645 observations of accesses to xls files, 1323 observations of accesses to ppt files, 857 observations of accesses to pdf files and 728 observations of accesses to "miscellaneous" files.

Examination for truncated Weibull distribution

The assumption of a mixed distribution function with a truncated Weibull distribution as continuous distribution will be examined first. The following zero hypothesis will be tested with the χ^2 adaptation test to significance level 0.001:

> H_0: The random variable X is descended from a main unit with the mixed distribution according to equation 2 with the truncated Weibull distribution function $F^*_{W>0}(x)$.

The zero hypothesis was not rejected in the case of observations of accesses to the xls files, the ppt files and the "miscellaneous" files (see Table 7). It must, in contrast, be rejected for doc and pdf files.

File type	$\hat{\alpha}$	$\hat{\beta}$	Rejection range	Result
doc	0.38	23.6	$T > \chi^2_{33;0.001} \Leftrightarrow 91.85 > 63.87$	H_0 rejected
xls	0.25	1.1	$T > \chi^2_{29;0.001} \Leftrightarrow 49.59 < 58.30$	H_0 not rejected
ppt	0.38	14.3	$T > \chi^2_{31;0.001} \Leftrightarrow 39.69 < 61.10$	H_0 not rejected
pdf	0.48	21.9	$T > \chi^2_{28;0.001} \Leftrightarrow 79.59 > 56.89$	H_0 rejected
miscellaneous	0.46	27.7	$T > \chi^2_{30;0.001} \Leftrightarrow 38.83 < 59.70$	H_0 not rejected

Table 7. χ^2 adaption tests on the sample divided according to file types for a mixed distribution function with truncated Weibull distribution.

The xls files, ppt files and miscellaneous files represent 53.89 % of all files in the sample. The mixed distribution function consisting of step function and truncated Weibull distribution function is a suitable distribution model for these file types. These conclusions can be applied directly to the generation of a migration rule in the framework of ILM.

Examination for truncated gamma distribution

The assumption of a mixed distribution function with the truncated gamma distribution as continuous distribution will be examined in this section. With the χ^2 adaption test, the zero hypothesis

> H_0: The random variable X is descended from a main unit with the mixed distribution according to equation 2 with the truncated gamma distribution function $F^*_{W>0}(x)$.

is tested to significance level 0.001.

Table 8 shows that the assumed distribution model with the truncated gamma distribution is suitable for modelling the random variable X: "Number of days since the last access" for ppt files and "miscellaneous" file types. The zero hypothesis must however be rejected for the other three file types.

File type	$\hat{\alpha}$	$\hat{\beta}$	Rejection range	Result
doc	0.21	279	$T > \chi^2_{32;0.001} \Leftrightarrow 71.31 > 62.49$	H_0 rejected
xls	0.10	141	$T > \chi^2_{25;0.001} \Leftrightarrow 172.59 > 52.62$	H_0 rejected
ppt	0.19	221	$T > \chi^2_{30;0.001} \Leftrightarrow 57.36 < 59.70$	H_0 not rejected
pdf	0.29	146	$T > \chi^2_{28;0.001} \Leftrightarrow 67.46 > 56.89$	H_0 rejected
miscellaneous	0.29	181	$T > \chi^2_{29;0.001} \Leftrightarrow 27.61 < 49.59$	H_0 not rejected

Table 8. χ^2 adaption tests on the sample divided according to file types for a mixed distribution function with truncated gamma distribution.

The result of section "Division of the sample according to file types" is that a suitable distribution model can be constructed for the xls files, ppt files and "miscellaneous" files contained in the sample: When a migration rule is generated for ILM, a mixed distribution function with a truncates Weibull distribution should be used for xls files, while for ppt files and "miscellaneous" file types a choice can be made between a Weibull or gamma distribution.

If a rule must be generated for doc and pdf files, the mixed distribution function with a truncated gamma distribution should be selected. In spite of negative test results, the gamma distribution is better suited for these cases than the Weibull distribution, since the former achieves lower test values during the χ^2 adaption test, which produces a better adaption.

Summary of the test results

It was established in section "Theoretical Distribution Models" that there is no suitable distribution model for the complete sample. For this reason, mixed distribution functions were introduced in section "Construction of a Mixed Distribution Function" and differently composed subgroups of the sample were examined for their distribution. Suitable distribution models were then successfully generated for some subgroups. Table 9 gives an overview of the test results.[2]

Criterion	Class	Distribution model
Age of the file	[0 days;365 days)	W(0.35,3.5)
	[365 days;730 days)	-
	[730 days;1772 days)	-
Number of accesses	[1 access;7 accesses)	-
	[7 accesses;15 accesses)	G(0.32,183)
	[15 accesses;292 accesses)	W(0.36,4.0)
File type	doc	-
	xls	W(0.25,1.1)
	ppt	W(0.38,14.3), G(0.19,221)
	pdf	-
	miscellaneous	W(0.46,27.7), G(0.29,181)

Table 9. Summary of the test results. W($\hat{\alpha}$, $\hat{\beta}$) = Weibull distribution, G($\hat{\alpha}$, $\hat{\beta}$) = gamma distribution, $\hat{\alpha}$ and $\hat{\beta}$ are the estimated values of the parameters. "-" = no suitable distribution model.

The criteria "Access type" and "File size" are not listed in the table, since the χ^2 adaption tests were negative in these cases.

As can be seen in Table 9, positive test results were achieved for some subgroups, e.g. xls files. These conclusions can be applied directly for the generation of migration rules (see section "Application of the Test Results").

It must still be clarified whether every file in the sample can be assigned to at least one subgroup, for which a distribution model exists. Two or three suitable distribution models have been determined for some files, since every file can be found in the samples divided according to "File age", according to "Number of accesses" and also according to "File type" (see table 10).

[2] "Distribution model" stands in the table for the suitable continuous and truncated distribution function in a mixed distribution function.

Number of accesses	File type				
	doc	xls	ppt	pdf	miscellaneous
[1 access;7 accesses)	-	○	○	-	○
[7 accesses;15 accesses)	●	○, ●	○, ●	●	○, ●
[15 accesses;292 accesses)	●	○, ●	○, ●	●	○, ●

Table 10. Files with minimum age of 365 days. "○": A suitable distribution model exists for this file type. "●": A suitable distribution model exists for files with this number of accesses. "-": A suitable distribution model does not exist for this file type.

Table 10 contains only files with a minimum age of 365 days, in order to show for which files a suitable distribution model does not exist. A suitable distribution model was established for all files with a lower age. A distribution model with positive test results cannot be specified for 129 of the 693 examined files. This applies to 98 doc and 31 pdf files with less than 7 accesses and an age of at least 365 days.

APPLICATION OF THE TEST RESULTS

We now have the method to valuate files for ILM. Some examples show the feasibility of the method:

Example1: Type: doc, Age: 50 days, Accesses: 10, last access 5 days ago.

This file has a probability of further access of 60.05 %.

Example 2: Type: pdf, Age: 420 days, Accesses: 3, last access 230 days ago.

This file has a probability of further access of 2.44 %.

Example 3: Type: other, Age: 30 days, Accesses: 20, last access 2 days ago.

This file has a probability of further access of 67.82 %.

Now migration rules for ILM will be derived.

In the considered case, F(x) is the distribution function of the random variable X: "Number of days since the last access". It can be calculated with the difference 1-F(x) = 1-P(X ≤ x) = P(X>x), with which probability p the "Number of days since the last access" exceeds a value x. Vice versa the number of days x, after which the file is accessed only with a certain probability p, can be calculated from the distribution function. The number of days x is the *threshold value* of the corresponding migration rule, which is as follows:

Migrate the file to the next lower level, if the probability of further accesses is below p=1%.

We apply this migration rule to the group of xls-files. The calculation will tell us, how long it takes until the access probability falls below 1%

It was shown in section "Examination for truncated Weibull-distribution" that the random variable X: "Number of days since the last access" can be modeled for xls files using a mixed distribution function consisting of a step function and a truncated Weibull distribution. The distribution function of the xls files $F_{xls}(x)$ results from the insertion of the following values in equation 2: The number of observations of accesses to xls files is 2645, of which 648 have zero days since the last access; the maximum number of days since the last access is 1050; the estimated values of the parameters are $\hat{\alpha} = 0.25$ and $\hat{\beta} = 1.1$ (see Table 7). The corresponding distribution function is then

$$F_{xls}(x) = \frac{648}{2645} \cdot \mathbf{1}_{[0,1050)}(x) + \frac{1997}{2645} F^{\star}_{W>0;xls}(x) \qquad (3)$$

where

$$F^{\star}_{W>0;xls}(x) = \frac{1 - e^{-(x/1,1)^{0,25}} - (1 - e^{-(6,94\cdot 10^{-4}/1,1)^{0,25}})}{1 - e^{-(1050/1,1)^{0,25}} - (1 - e^{-(6,94\cdot 10^{-4}/1,1)^{0,25}})} \qquad 6,94 \cdot 10^{-4} \leq x \leq 1050$$

$F^*_{W>0;xls}(x)$ is the Weibull distribution function truncated to the interval $[6.94 \cdot 10^{-4};1050]$.

In this example, an xls file should be migrated only if further accesses are to be expected only with a probability of 1 %. The following equation must be solved to determine after how many days this will be the case:

$$0,99 \stackrel{!}{=} F_{xls}(x) \tag{4}$$

It results x = 630.84.

The resulting migration rule is:

Migrate xls-files to the next lower level, if 630,48 days have expired since the last access.

One can generate migration rules for other subgroups of the sample in the same manner, if a suitable distribution model exists (see Tables 9 and 10).

SUMMARY AND OUTLOOK

Proper information valuation is the first step towards ILM automation. Existing valuation methods either use metadata or look at the history and generate a value in terms of "amount of dollars" or "a decimal figure within an interval".

We presented a case study and derived distribution functions to describe the access behaviour of different MS Office file types. We demonstrated that the future access of a file can be predicted from observed access data and that this can be used as a metric for file valuation.

The value of a file is its percentage of further accesses. This is a new way of valuation. The advantages are that it is simple, does not need metadata and fits to ILM automation. The disadvantage of this method is that it does not consider legislation. In general legislation and compliance are important issues in ILM. Therefore actions resulting from this method should be aligned with legislation guidelines.

Starting from this work, ILM rules can be generated directly. Therefore our next step is to implement a simulator. This will provide the possibility to simulate ILM scenarios and to check the rules for quality.

REFERENCES

1. Beigi, M., Devarakonda, M., Jain, R., Kaplan, M., Pease, D., Rubas, J., Sharma, U., Verma, A., (2005) "Policy-Based Information Lifecycle Management in a Large-Scale File System". In Proceedings of the Sixth IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY'05) pp. 139-148
2. Chen, Y. (2005) Information valuation for Information Lifecycle Management. In: Proceedings of the Second International Conference on Autonomic Computing (ICAC'05), pages 135-146.
3. Denning, P. J., (1968) "The working set model for program behavior". Communications of the ACM, 11(5), 1968.
4. Denning, P. J., (1980) "Working sets past and present". IEEE Transactions on Software Engineering, SE-6(1):64–84, 1980.
5. Effelsberg, W. and Haerder, T., (1984) "Principles of Database Buffer Management". ACM Transactions on Database Systems, 9(4), 1984.
6. Gibson, T. and Miller, E. (1999), An Improved Long-term File-Usage Prediction Algorithm, University of Maryland, Baltimore County, USA.
7. Gibson, T., Miller, E. and Long, D. (1998), Long-term File Activity and Inter-Reference Patterns, 24th International Conference on Technology Management and Performance Evaluation of Enterprise-Wide Information Systems, Computer Measurement Group, Anaheim, California, USA.
8. Hartung, J. (1995), Statistik - Lehr- und Handbuch der angewandten Statistik, 10. Aufl., München u.a., Oldenbourg.
9. Lawrie, D. H. Randal, J. M. and. Barton, R. R., (1982) "Experiments with Automatic File Migration," IEEE Computer 15(7), 1982, pp. 45-55.

10. Page, L., Brin, S., Motwani, R. and Winograd. T., (1999) "The pagerank citation ranking: Bringing order to the web". http://dbpubs.stanford.edu:8090/pub/1999-66, 1999.
11. Peterson, M (2004)., ILM Definition and Scope - An ILM Framework, SNIA Data Management Forum, Version 2.3, July 2004. www.snia.org/tech_activities/dmf/docs
12. Reiner, D. Press, G., Lenaghan, M., Barta, D., Urmston, R., (2004) "Information Lifecycle Management: The EMC Perspective". Proceedings of the 20th International Conference on Data Engineering (ICDE'04) 1063-6382/04
13. Ridings, R. and Shishigin, M., (2002) "PageRank Uncovered". http://www.voelspriet2.nl/PageRank.pdf, 2002.
14. Sachs, L. (2004), Angewandte Statistik - Anwendung statistischer Methoden, 11. Aufl., Berlin u.a., Springer.
15. Schlittgen, R. (2003), Einführung in die Statistik - Analyse und Modellierung von Daten, 10. Aufl., München u.a., Oldenbourg.
16. Schmitz, C. (2004), Entwicklung einer optimalen Migrationsstrategie für ein hierarchisches Datenmanagement System, Forschungszentrum Jülich GmbH, Jülich.
17. Smith, A. J., (1982) "Long Term File Migration: Development and Evaluation of Algorithms," Communications of ACM 24(8), 1982, pp.521-532.
18. Strange, S. (1992), Analysis of Long-term Unix File Access Patterns for Application to Automatic File Migration Strategies, University of California, Berkeley, California, USA.
19. Tanaka, T., Ueda, R., Aizono, T., Ushijima, K., Naitih, I., Komoda, N. (2005) Proposal and Evaluation of Policy Description for Information Lifecycle Management Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05) 0-7695-2504-0/05
20. Turczyk, L., Berbner, R., Heckmann, O., Steinmetz, R., (2006) "A formal approach to Information Lifecycle Management". In Proceedings of 17th Annual IRMA International Conference, Washington D.C.