World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:6, 2015

An Automatic Bayesian Classification System for File Format Selection

Roman Graf, Sergiu Gordea, Heather M. Ryan

Abstract—This paper presents an approach for the classification of an unstructured format description for identification of file formats. The main contribution of this work is the employment of data mining techniques to support file format selection with just the unstructured text description that comprises the most important format features for a particular organisation. Subsequently, the file format indentification method employs file format classifier and associated configurations to support digital preservation experts with an estimation of required file format. Our goal is to make use of a format specification knowledge base aggregated from a different Web sources in order to select file format for a particular institution. Using the naive Bayes method, the decision support system recommends to an expert, the file format for his institution. The proposed methods facilitate the selection of file format and the quality of a digital preservation process. The presented approach is meant to facilitate decision making for the preservation of digital content in libraries and archives using domain expert knowledge and specifications of file formats. To facilitate decision-making, the aggregated information about the file formats is presented as a file format vocabulary that comprises most common terms that are characteristic for all researched formats. The goal is to suggest a particular file format based on this vocabulary for analysis by an expert. The sample file format calculation and the calculation results including probabilities are presented in the evaluation section.

Keywords—data mining, digital libraries, digital preservation, file format.

I. Introduction

IN recent years, libraries, archives and museums have created new digital collections that comprise millions of objects, and the goal is to make them available on a long term basis. One of the core preservation activities deals with the evaluation of appropriate formats used for encoding digital content. The preservation risks for a particular file format are difficult to estimate as described in [5]. The definition of risk factors and associated metrics is still an open research topic in the digital preservation community. Involvement of digital preservation experts is required for collecting complete information and evaluating preservation risks as shown in [1]. Currently, each institution selects its own file formats for long term preservation depending on the particular project, preservation goals, workflows and assets. Due to the scale of digital information that has to be managed, memory institutions are facing challenges regarding preservation, maintenance, and quality assurance of these collections. For that reason, automated solutions for data management and digital preservation are absolutely necessary. Many file formats are properly documented, are open-source and well supported by software vendors. Other formats may be outdated or no longer functional with modern software or hardware. There are also custom/proprietary formats - which may be obsolete and not renderable with commodity hardware. To address these problems, we employ the File Format Metadata Aggregator (FFMA) system [4] and the information integration approach. FFMA is a part of a knowledge base recommender DiPRec as shown in [3], which reuses the experience of building preservation planning tools and offers assessment for long-term preservation of digital content. This tool performs an analysis of file formats based on the concept of risk scores. But FFMA does not provide information about format specification that could be very useful for analysis since open repositories include only file format descriptions. Therefore, the file format specifications aggregation and analysis is an important open issue. The proposed approach facilitates the selection of institutional file formats. The specifications knowledge base is aggregated from different open sources on the Web like fileformat.info, wikipedia etc. Collected information is not structured and not homogenous. Every format is documented in a different way. The vendors of proprietary formats do not provide specifications, but standardisation and homogenity of specifications is not required for our approach since we analyse unstructured text. The goal of this approach is to help select an institutional file format. Fig. 1 shows the general workflow for the selection of an institutional file format. The data for file format calculation is aggregated from the classification knowledge base. The knowledge base employs the aggregated specifications for the computation of the institutional file format vocabulary. Classification of the unstructured file format descriptions based on Bayes theorem provides estimation about the best matching file format. The novelty of this technical solution is the employment of data mining methods to facilitate complex information research on file formats for preservation experts. Decision support based on this calculation approach and expert knowledge base is designed to support institutions like libraries and archives with assessment for analyzing their digital assets. This paper is structured as follows: Section II gives an overview of related work and concepts. Section III explains the file format selection workflow and also covers data mining issues. Section IV presents the experimental setup, applied methods and results. Section V concludes the paper and provides an outlook on planned future work.

Roman Graf and Sergiu Gordea are with the Department Digital Safety & Security, Austrian Institute of Technology GmbH, Austria, (e-mail: {roman.graf,sergiu.gordea}@ait.ac.at).

Heather M. Ryan is with Library & Information Science Program, University of Denver (e-mail: heather.m.ryan@du.edu).

II. Related Work

The research on risk management in digital collections increasingly gains in importance. It is difficult to guarantee the longivity of digital information. The investigation [10] aims at risk assessment of migrating of file formats. Accurate format identification and rendering is a challenging task due to malformed MIME types, rendering expenses, dependence on content not embedded in the file, missing colour tables, changed fonts, etc. In [9], the author examines how the network effects could stabilise formats against obsolescence. The result of evaluation demonstrates that most formats last much longer than five years, that network effects stabilize formats, and that new formats appear at a modest, manageable rate. However, a number of formats are fading from use and every corpus contains its own biases. Digital preservation tools like PANIC [8], AONS II [11], SPOT [12], P2 registry [2], aimed at identifying file formats used for encoding digital collections and informing repository managers of events that might impact access to the stored content. They also define alerting mechanisms when file formats become obsolete. As distinct from our approach they do not apply expert knowledge and do not specify risk factors that may influence file format endangerment. The FFMA [3] is a preservation planning tool that offers assessment for long-term preservation of digital content. This tool performs an analysis of file formats based on the concept of risk scores. Selected institutional risk profile in conjunction with FFMA can calculate endangerment risks for selected file format. There are multiple influential algorithms [13] (k-Means, SVM, kNN, Naive Bayes), which can be applyied in data mining. The Naive Bayes algorithm is very good at the matching for classification task in our approach. Bayesian networks [7] extended with statistical techniques are used in data mining to encode probabalistic relationships among variables of interest. Such networks combine prior uncertain expert knowledge with the data and are related to graphical modelling techniques for supervised and unsupervised learning and for learning with incomplete data. In our approach we do not use rule bases, decision trees or artificial neural networks but employ the Naive Bayes method for probabilities calculation. In the proposed approach we intend to apply standard statistics and data mining methods for digital preservation tasks. The proposed system is unique for the given domain.

III. File Format Selection Method

In the presented approach the specification knowledge base is aggregated by the domain experts and it consists of multiple text files for each of 12 exemplarly selected file formats. This selection is based on the format selection from [6]. These formats are “BMP”, “DOC”, “DXF”, “GIF”, “HTML”, “JPG”, “MP3”, “PDF”, “PNG”, “PPT”, “SXW” and “TIF”. Since the “MAC” format is a proprietary format it is difficult to obtain its specification and we amended it in our analysis. The file format selection is conducted according the workow shown in Fig. 1.

The workflow execution starts with the aggregation of format specifictaions. The specifications data manually aggregated by domain experts is stored in a text file. The data is arranged in folders accordingly to a particular file format. Each folder comprises specifications from different sources for a particular format. Each folder is a format category for our calculations. In the second step the workflow execution proceeds with the reading of the aggregated data from the files and computation of the format vocabulary. In this step we count the words in the specifications grouped by category and add them to the format vocabulary. The words from the ignore list are removed before the format vocabulary calculation starts. Additionally, vocabulary words should occure in specifications at least a number of times. This number is defined by an associated threshold value W_{min} (1).

Having the automatically calculated format vocabulary limited by the thouthands of the most common words from the specifications at hand, we can train our classifier and compute probabilities for each word in the format vocabulary taking in account the category, the word occurencies number and the total number of words in the vocabulary.

The classification of the unstructured format description occures in the fourth step. Having an input text in the form of

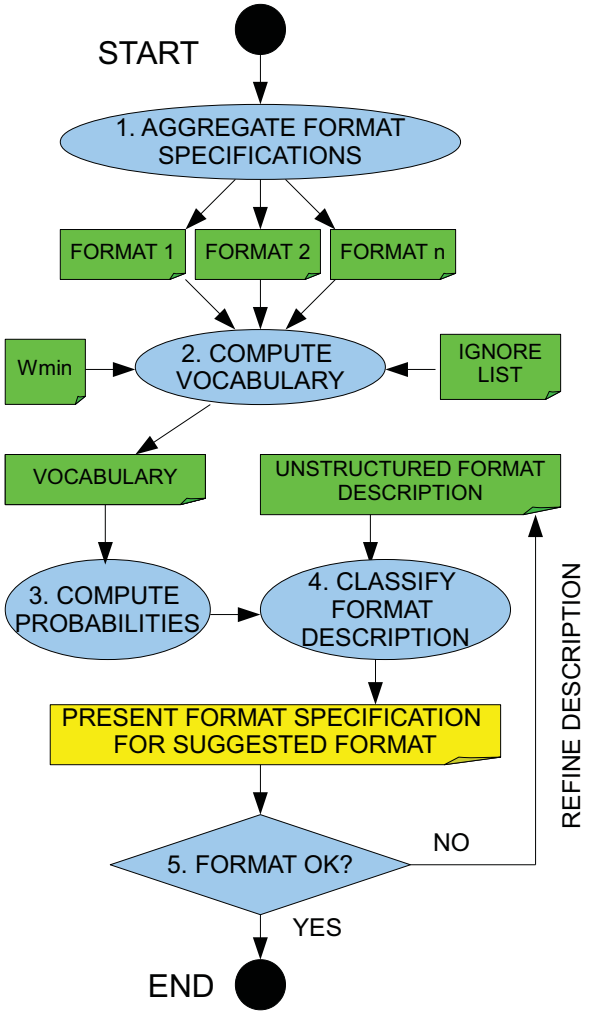

Fig. 1: The file format selection workflow.

an unstructured format description we can classify the given format description to one of the format specification category employing the naive Bayes algorithm [15], [14] (1) in the third step. This Formula shows the probability of the format description D (2) belonging to the file format category c. The probability of the format description D is a product of all specifications fs that are comprised in the format vocabulary.

$$p\left(c|D\right) = \frac{p\left(D|c\right)p\left(c\right)}{p\left(D\right)}. \quad (1)$$

$$D = \left(fs_1, fs_2 ... fs_{12}\right). \quad (2)$$

Since the computed probabilities are very small we employ the log function and add a log of probabilities. The naive Bayes algorithm picks the file format category with the highest probability. The selected format should be validated by an expert in the last step of the workflow. In the case that institutional expert is not satisfied with this format the format description should be refined with additional key words that are important for institution and workflow returns to the fourth step. Additionally, new format specifications can be added to the knowledge base. In this case the whole process should begin from the first step since the training phase is necessary for classification step. The possible hypotheses for file format selection calculation are that an institutional format description belongs to one of the aggregated format specifications. Therefore, a format description that matches a particular format vocabulary words is more likely to have an associated file format specification for selection and decision making for the digital preservation long-term planning. For Bayesian approach we assume that words in specifications are independent.

IV. Evaluation

The goal of this evaluation was the leveraging of the domain expert knowledge base for selection of the file format as described in the workflow for file format selection (see Fig. 1).

A. Hypothesis and Evaluation Methods of the File Format Selection

The hypothesis is that format specifications aggregated from a domain expert knowledge base can be selected by the unstructured textual format description. Therefore, a human expert can easily select a file format with particular features specified in text form for specific digital preservation task. Our approach should give an organisation a base of information that helps to select between alternative file formats with the required feature set. This decision should be the best choice for the organisation's preservation programme. The employment of data mining techniques facilitates this task for a human expert by performing complex calculations and comparisons.

In three evaluation scenarios, we performed the sample file format selection calculation. The hypothesis is that an institutional expert will define some of the most important file format characteristics in a text file and input them into the data mining tool. The output of the tool should be the matching file format specification from the knowledge base. Thus, a preservation expert can adjust the required file format characteristics in order to select the best matching format and to reduce preservation risks. The differences between the three scenarios are that in the first scenario we do not employ a list with words that should be ignored by the format vocabulary calculation. In the second scenario we apply such a list with the 175 most often used english words like "the", "a", "do", "could" etc. These words do not provide semantic value for the given analysis, reduce performance, and increase complexity. In the third scenario we extend the ignore list by the 75 words specific for file format specifications, which are not important for the format description. These are numbers "1", "2" or special characters like "=", "#" or "*".

Evaluation took place on an Intel Core i73520M 2.66GHz computer using Python 2.7 language on Windows OS. We evaluated different short file format descriptions taken from the Web[1] and calculation time for each evaluation scenario.

B. Evaluation Data Set

The basis for the file format selection calculation was provided through a format specification research in which digital preservation experts aggregated format specifications for 12 selected file formats. For the knowledge base we aggregated format specifications from different Web sources for "BMP", "DOC", "DXF", "GIF", "HTML", "JPG", "MP3", "PDF", "PNG", "PPT", "SXW" and "TIF".

C. Experimental Results and Its Interpretation

The experimental results are presented in three tables. Table I demonstrates the classification results for given format descriptions without employing the ignore list. The first column "Format" in the Table I shows the expected file formats for the given associated format description. Having the specification knowledge base and a classifier from the preevaluation step we use it as a basis for the format selection. The second column "Correct" provides calculation results for each format. "1" stands for the correct selection and "0" for negative result. The correct selection means that an input unstructured short format description indeed has the highest match with the associated file format specification. The following columns represent calculated probabilities, whereas the most nearest probability starts from the column with index "1" and the lowest probability is presented in the column with index "12". Each probability cell comprises the file format category e.g. "BMP" and associated negative log value. The smaller the log value, the higher the probability that the associated file format category belongs to the expected format specification.

The calculation time for this scenario is 0,786 seconds, accuracy is 75% and vocabulary size 6144 words.

The calculation time for the second scenario (Table II) is 0,765 seconds, accuracy is 83% and vocabulary size 6119 words.

[1] http://fileinfo.com/

TABLE I: The Classification Results without Ignore List.

Format	Correct	1	2	3	4	5	6	7	8	9	10	11	12
BMP	1	BMP/401	GIF/428	TIF/430	PNG/433	DXF/465	PPT/475	HTML/476	MP3/478	PDF/479	JPG/483	DOC/494	SXW/507
DOC	1	DOC/278	PPT/285	TIF/287	BMP/294	PNG/301	PDF/301	DXF/304	GIF/306	HTML/308	MP3/316	JPG/323	SXW/326
DXF	1	DXF/487	TIF/496	PPT/496	GIF/505	BMP/508	PNG/510	DOC/513	PDF/514	MP3/517	JPG/535	HTML/546	SXW/559
GIF	1	GIF/829	PNG/852	BMP/855	TIF/860	PPT/897	HTML/902	PDF/905	DXF/917	MP3/923	JPG/927	DOC/942	SXW/1000
HTML	1	HTML/575	GIF/579	DXF/583	PDF/585	PNG/588	PPT/591	TIF/592	BMP/603	DOC/606	MP3/607	JPG/612	SXW/661
JPG	0	TIF/465	BMP/476	PNG/481	GIF/482	JPG/491	MP3/496	PPT/499	PDF/506	DOC/516	DXF/519	SXW/536	HTML/542
MP3	1	MP3/668	BMP/720	TIF/722	PPT/723	PNG/727	GIF/734	PDF/737	DXF/738	DOC/745	JPG/768	SXW/789	HTML/810
PDF	1	PDF/956	TIF/978	PPT/978	BMP/1001	GIF/1003	PNG/1004	DOC/1005	DXF/1014	MP3/1020	HTML/1034	JPG/1071	SXW/1077
PNG	0	GIF/840	PNG/848	BMP/869	TIF/870	PPT/901	PDF/903	MP3/942	DOC/944	JPG/944	DXF/947	HTML/954	SXW/985
PPT	1	PPT/264	PNG/283	PDF/285	DOC/285	TIF/285	DXF/283	BMP/288	GIF/289	MP3/291	HTML/299	SXW/305	JPG/309
SXW	0	DOC/270	PPT/274	TIF/282	PNG/282	PDF/284	SXW/285	BMP/291	GIF/295	MP3/296	DXF/298	JPG/309	HTML/314
TIF	1	TIF/328	BMP/333	PNG/334	GIF/335	PPT/343	DXF/346	PDF/349	DOC/353	MP3/354	JPG/354	HTML/358	SXW/373

TABLE II: The Classification Results for Ignore List with 175 English Words.

Format	Correct	1	2	3	4	5	6	7	8	9	10	11	12
BMP	1	BMP/305	GIF/332	TIF/335	PNG/336	DXF/368	PPT/376	PDF/378	SXW/380	HTML/381	MP3/383	JPG/386	DOC/392
DOC	1	DOC/216	PPT/224	TIF/226	BMP/234	PDF/238	PNG/240	DXF/245	GIF/246	SXW/246	HTML/248	MP3/255	JPG/264
DXF	1	DXF/291	PPT/301	TIF/305	BMP/309	PDF/310	DOC/310	PNG/310	SXW/313	GIF/313	MP3/319	JPG/333	HTML/356
GIF	1	GIF/525	PNG/534	BMP/543	TIF/554	PPT/579	PDF/583	SXW/598	HTML/600	DXF/603	JPG/606	DOC/609	MP3/614
HTML	0	PDF/366	HTML/372	GIF/374	PPT/376	PNG/376	DXF/379	SXW/383	DOC/383	TIF/388	BMP/381	MP3/397	JPG/403
JPG	0	TIF/354	BMP/365	PNG/369	GIF/372	JPG/381	PPT/383	MP3/384	PDF/390	SXW/391	DOC/399	DXF/408	HTML/430
MP3	1	MP3/482	BMP/531	PPT/533	TIF/535	PNG/537	PDF/541	SXW/543	GIF/546	DOC/548	DXF/549	JPG/577	HTML/625
PDF	1	PDF/644	PPT/677	TIF/681	DOC/691	SXW/693	PNG/694	BMP/698	GIF/702	DXF/712	MP3/718	HTML/742	JPG/765
PNG	1	PNG/582	GIF/585	BMP/608	TIF/610	PDF/634	PPT/636	SXW/658	DOC/669	JPG/678	MP3/683	DXF/685	HTML/699
PPT	1	PPT/200	PNG/217	PDF/219	DOC/219	TIF/220	BMP/222	DXF/223	SXW/224	GIF/225	MP3/225	HTML/234	JPG/243
SXW	1	SXW/210	DOC/210	PPT/216	PDF/224	TIF/225	PNG/225	BMP/233	MP3/238	GIF/238	DXF/241	JPG/251	HTML/258
TIF	1	TIF/259	PNG/265	BMP/265	GIF/267	PPT/274	DXF/277	PDF/278	SXW/282	DOC/283	MP3/287	JPG/287	HTML/290

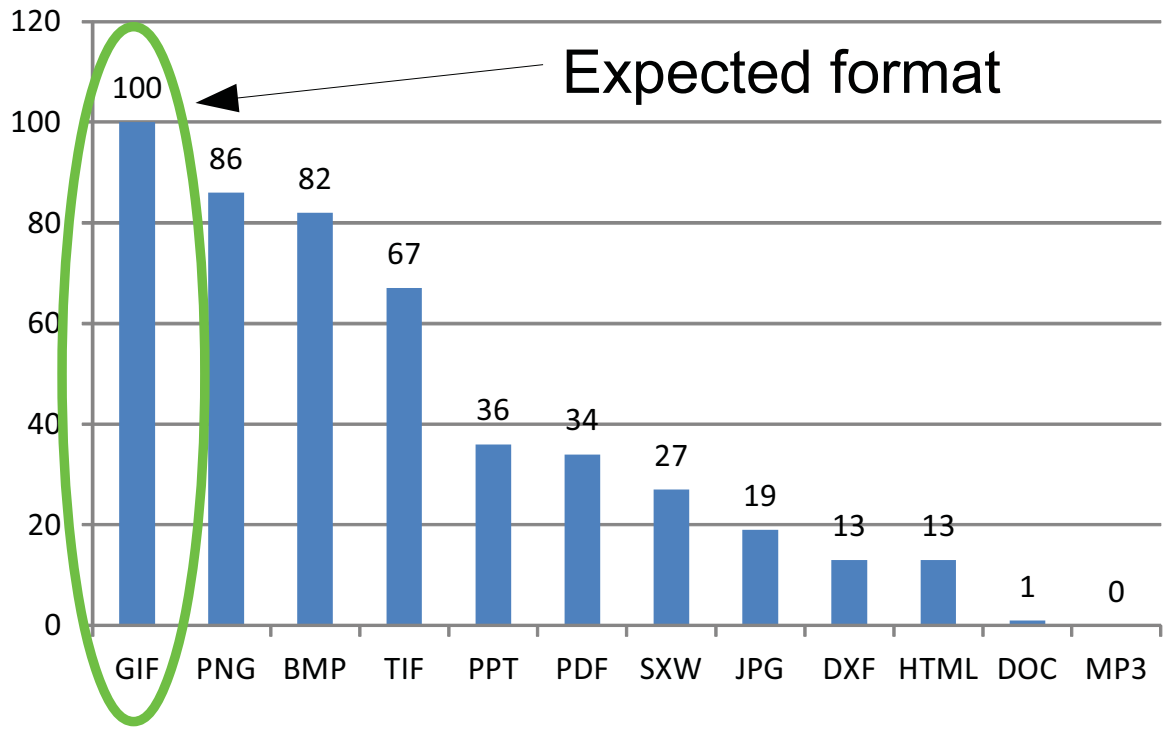

Fig. 2: Diagram for correct detection and relation between probabilities of analyzed format categories.

The calculation time for the third scenario (Table III) is 0,753 seconds, accuracy is 83% and vocabulary size 5948 words.

The vocabulary is a dictionary that contains pairs. The key in the pair is a word and value is a count of this word in the specifications. E.g. in the third scenario “data”=1291, “image”=1109, “attribute”=933, “element”=909, “table”=900, “file”=869.

These results demonstrate that employing the ignore lists improves performance and calculation accuracy. The more accurate the word choice in ignore lists, the better results come out. For more exact accuracy calculation we need more file format specifications and better ignore lists.

$$V_i = \frac{(V_{imax} - C_i) * 100}{(V_{imax} - V_{imin})}, i = 1..12. \quad (3)$$

TABLE III: The Classification Results for Ignore List with 175 English Words and 75 Specific Ignore Words for Formats.

Format	Correct	1	2	3	4	5	6	7	8	9	10	11	12
BMP	1	BMP/270	GIF/296	TIF/299	PNG/302	DXF/328	SXW/338	PDF/339	PPT/340	HTML/343	MP3/345	JPG/346	DOC/354
DOC	1	DOC/201	PPT/210	TIF/211	BMP/218	PDF/222	PNG/225	GIF/229	SXW/230	DXF/231	HTML/233	MP3/240	JPG/247
DXF	1	DXF/277	PPT/288	TIF/211	BMP/218	PNG/295	SXW/296	DOC/296	PDF/297	GIF/298	MP3/304	JPG/214	HTML/338
GIF	1	GIF/468	PNG/480	BMP/484	TIF/497	PPT/524	PDF/526	SXW/532	JPG/539	DXF/545	HTML/545	DOC/551	MP3/556
HTML	0	PDF/352	HTML/354	GIF/358	PNG/361	PPT/363	DXF/365	SXW/366	DOC/369	TIF/372	BMP/381	MP3/382	JPG/383
JPG	0	TIF/345	BMP/355	PNG/361	GIF/361	JPG/368	MP3/375	PPT/376	PDF/382	SXW/382	DOC/391	DXF/398	HTML/419
MP3	1	MP3/447	BMP/493	PPT/499	TIF/499	PNG/501	SXW/501	PDF/505	GIF/510	DXF/512	DOC/513	JPG/535	HTML/586
PDF	1	PDF/602	TIF/636	PPT/638	SXW/644	PNG/649	DOC/651	BMP/653	GIF/658	DXF/669	MP3/674	HTML/694	JPG/712
PNG	1	PNG/566	GIF/566	BMP/589	TIF/592	PDF/619	PPT/622	SXW/641	DOC/654	JPG/654	MP3/667	DXF/667	HTML/677
PPT	1	PPT/180	DOC/197	PDF/198	PNG/198	TIF/199	SXW/199	BMP/200	DXF/200	GIF/204	MP3/205	HTML/216	JPG/220
SXW	1	SXW/210	DOC/210	PPT/215	TIF/223	PDF/223	PNG/224	BMP/231	GIF/235	MP3/237	DXF/240	JPG/248	HTML/254
TIF	1	TIF/250	PNG/256	BMP/256	GIF/258	PPT/267	DXF/269	PDF/270	SXW/273	JPG/276	DOC/276	MP3/279	HTML/279

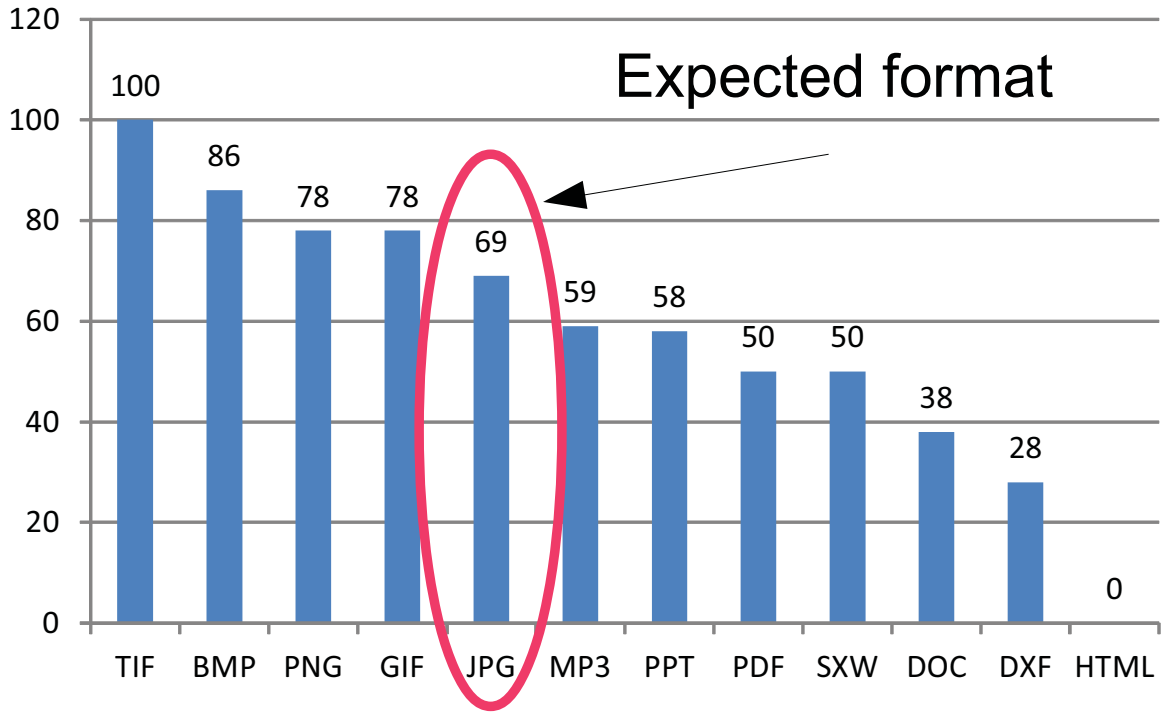

Fig. 3: Diagram for incorrect detection and relation between probabilities of analyzed format categories.

For evaluation of the selection correctness in the third scenario two selected probability profiles are used: “GIF” (see Table III and Fig. 2) and “JPG” (see Table III 3). Formula 3 is employed for presentation of probability profile. The V_i is a normalized probability value in the table row e.g. “GIF” row. i is a current column number standing for associated file format. The V_{imax} is the maximal probability value in the row. The V_{imin} is the minimal probability value in the row. The C_i is the current calculated log probability value in the row.

The computation by means of naive Bayes algorithm for the given input “GIF” format description returns the file format category “GIF” (see Fig. 2 and “GIF” row in the Table III). That means that at the beginning of evaluation, institutional format characteristics are most likely belong to the “GIF” format specification. Having this information, institutional experts can analyse this specification and if necessary, adjust the format description in order to change the selected file format. E.g. if an expert expects an image format he/she would like to have words like “image”, “colour”, “RGB” in his/her format description.

The computation for the given input “JPG” format description returns the file format category “TIF” (see Fig. 3 and “JPG” row in the Table III). That means that at the beginning of evaluation, institutional format characteristics most likely belong to the “TIF” format specification and not as expected to “JPG”. The reason is that the input format description is relatively short and by searching for image formats, multiple file formats can be described by the same words. In this case the expected “JPG” is only on the fifth place preceded by another image formats “TIF”, “BMP”, “PNG” and “GIF”. Addition of more specific words that are characteristic only for “JPG” format in the input text would improve the accuracy of selection.

This approach should support the definition of institutional policies for preservation file format selection. The knowledge about formats reduces the endangerment level of a digital collection. Employing the provided algorithm the institutional expert can find the file format that is most similar to its short

textual description.

These results (accuracy 83,33%) demonstrate (see Fig. 2 and Tables II, III) that a semi-automatic approach for file format selection is very effective and it is a signicant improvement compared to manual analysis.

V. Conclusion

In this work we presented an approach for the classification of an unstructured format description for identification of file formats.

The main contribution of this work is the employment of data mining techniques to support file format selection with just the unstructured text description that comprises the most important format features for a particular organisation. The resulting Bayesian probability is used to support digital preservation experts with semi-automatic estimation of required file format.

The presented method employs a format specification knowledge base aggregated from web sources in order to select file format for particular institution.

To facilitate easier format selection, the aggregated information about the file formats is presented as a file format vocabulary. The proposed methods improve the usability of file format specifications and the quality of a digital preservation process.

We make use of data mining techniques like the naive bayes method in order to analyse aggregated data. The employment of the naive Bayes algorithms classifies the unstructured format description and makes recommendation to an expert regarding the most highly matching file format.

In the evaluation section, different configurations for file format calculation are exposed. Using the developed approach and adjusting input data, specification amount and ignore list selection, experts have the ability to choose the appropriate file format for digital preservation planning in their institution.

The presented approach is meant to facilitate decision making with regard to preservation of digital content in libraries and archives using domain expert knowledge. As future work we plan to extend the specifications knowledge base and to increase the amount and quality of aggregated expert information.

References

[1] P. Ayris, R. Davies, R. McLeod, R. Miao, H. Shenton, and P. Wheatley. The life2 final project report. Final project report, LIFE Project, London, UK, 2008.

[2] L. C. David Tarrant, Steve Hitchcock. Where the semantic web and web 2.0 meet format risk management: P2 registry. *International Journal of Digital Curation*, 6(1):165–182, 2011.

[3] S. Gordea, A. Lindley, and R. Graf. Computing recommendations for long term data accessibility basing on open knowledge and linked data. *Joint proceedings of the RecSys 2011 Workshops Decisions@RecSys'11 and UCERSTI 2*, 811:51–58, November 2011.

[4] R. Graf and S. Gordea. Aggregating a knowledge base of file formats from linked open data. *Proceedings of the 9th International Conference on Preservation of Digital Objects*, poster:292–293, October 2012.

[5] R. Graf and S. Gordea. A risk analysis of file formats for preservation planning. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres2013)*, pages 177–186, Lissabon, Portugal, Sep 2013. Biblioteca Nacional de Portugal, Lisboa.

[6] R. Graf, S. Gordea, and H. Ryan. A model for format endangerment analysis using fuzzy logic. In *Proceedings of the 11th International Conference on Digital Preservation (iPres2014)*, pages 160–168, Melbourne, Australia, Oct 2014. State Library of Victoria, Melbourne.

[7] D. Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1(1):79–119, 1997.

[8] J. Hunter and S. Choudhury. Panic: an integrated approach to the preservation of composite digital objects using semantic web services. *International Journal on Digital Libraries*, 6, (2):174–183, September 2006.

[9] A. N. Jackson. Formats over time: Exploring uk web history. *Proceedings of the 9th International Conference on Preservation of Digital Objects*, pages 155–158, October 2012.

[10] G. W. Lawrence, W. R. Kehoe, O. Y. Rieger, W. H. Walters, and A. R. Kenney. Risk management of digital information: A file format investigation. june 2000.

[11] D. Pearson and C. Webb. Defining file format obsolescence: A risky journey. *The International Journal of Digital Curation*, Vol 3, No 1:89–106, July 2008.

[12] S. Vermaaten, B. Lavoie, and P. Caplan. Identifying threats to successful digital preservation: the spot model rsik assessment. *D-Lib Magazine*, 18(9/10), September 2012.

[13] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[14] R. Zacharski. *A Programmer's Guide to Data Mining: The Ancient Art of the Numerati*. 2012.

[15] H. Zhang. The Optimality of Naive Bayes. In V. Barr and Z. Markov, editors, *FLAIRS Conference*. AAAI Press, 2004.